Statistical Practice and Reproducibility in Behavioural Science

Kenneth Teck Kiat Lim

A thesis submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

Behavioural Science Group and Department of Statistics

The University of Warwick

September 2019

Contents

Acknowledgements **ix**

Declaration **xi**

Abstract **xiii**

1 Introduction **1**

2 Statistical Practice and Replication Success in Behavioural Science: An Evaluation of Published Studies **3**

Abstract 3

2.1 Introduction 4

2.2 Developing a Checklist 7

2.3 Pilot Evaluation of Statistical Practice 9

2.4 Main study: Independent Evaluations of Statistical Practice 11

2.5 Discussion 29

2.6 Directions for Future Research 41

2.7 Conclusion 42

3 The Sign Effect in Intertemporal Choice **45**

3.1 Introduction 45

3.2 The sign effect anomaly 48

3.3 Formalising the sign effect with probabilities 56

4 A Systematic Review of the Sign Effect Anomaly 59

4.1 Study aim . . . 59

4.2 Data source . . . 60

4.3 Descriptive statistics at the question level . . . 63

4.4 Sign effect analysis . . . 72

4.5 Discussion . . . 97

5 Statistical modelling of individual participant responses to monetary gains and losses 101

5.1 Faralla et al. (2012) . . . 102

5.2 Xu et al. 2009 . . . 124

5.3 Han and Takahashi (2012) . . . 141

5.4 Hardisty et al. (2013) . . . 154

5.5 Hardisty and Weber (2009) Experiment 1 . . . 164

5.6 Hardisty and Weber (2009) Experiment 2 . . . 180

5.7 Discussion . . . 192

6 Conclusion 199

7 Appendix 203

7.1 Search terms . . . 203

7.2 The 55 checklist items used to calculate the score . . . 205

7.3 Checklist items not included in the calculation of score . . . 208

7.4 Checklist results by replication success categories . . . 210

Bibliography 225

List of Tables

2.1 Example of how guidelines were rephrased for the checklist. 8

2.2 Table of 58 checklist items with 'Yes' or 'No' responses. 15

3.1 The opportunity cost approach to understand the sign effect using questions 1a and 1b as an example. 49

3.2 Illustration of responses to a question pair in a 2 × 2 table using an opportunity cost approach. 49

3.3 Individual participant responses to a question pair using a discounting approach. 50

4.1 Summary of characteristics for the 9 included studies from 7 papers. . 65

4.2 Summary of key characteristics related to the sign effect for the 9 included studies 74

4.3 Percentage of switch points for gains and losses across all participants and trials. 83

4.4 Number of question pairs presented to each participant for each unique combination of factors based on the study design. 85

4.5 Percentage of switch points for gains and losses across all participants and trials. 87

4.6 Modelling individual participant responses to a question pair using a discounting approach. 89

4.7 Proportion of sooner gains and later losses, and the sign effect size. . 90

4.8 Summary of the sign effect and discounting behaviour between studies based on individual participant data 91

4.9 Illustration of responses to a question pair in a 2 x 2 table. 93

4.10 Breakdown of number and percentage of participants who did not discount any gain or any loss. 95

5.1 Number of questions for each unique value of amount ratio rounded to two decimal places. 107

5.2 Percentage of later choices across all participants for each unique value of amount ratio. 109

5.3 Table of logistic regression results from attempting to replicate Faralla et al. (2012). 116

5.4 Table of multilevel logistic regression results. 119

5.5 Diagnostics results for model 5. The later choice is taken as a "positive", while the sooner choice, a "negative". 123

5.6 Breakdown of the number of questions by the different time delays. . 126

5.7 Number of questions for each unique value of amount ratio rounded to two decimal places. 130

5.8 Number of questions and duplicated values of amount difference for each unique factorial combination. 132

5.9 Table of multilevel logistic regression results. 137

5.10 Diagnostics results for model 4. The later choice is taken as a "positive", while the sooner choice, a "negative". 140

5.11 Table of multilevel logistic regression results. 149

5.12 Diagnostics results for model 3. The later choice is taken as a "positive", while the sooner choice, a "negative". 152

5.13 Table of multilevel logistic regression results. 160

5.14 Diagnostics results for model 4. The later choice is taken as a "positive", while the sooner choice, a "negative". 163

5.15 Characteristics of participants by gender. 165

5.16 Table of multilevel logistic regression results. 176

5.17 Diagnostics results for model 2. The later choice is taken as a "positive", while the sooner choice, a "negative". 178

5.18 Table of multilevel logistic regression results. 188

5.19 Diagnostics results for model 2. The later choice is taken as a "positive", while the sooner choice, a "negative". 191

7.1 Responses to checklist items by replication success categories. 211

List of Figures

2.1 Sampling frame for a pilot evaluation of studies. 10

2.2 Scores of 39 behavioural science studies evaluated independently, by journal. 14

2.3 Scores of 39 behavioural science studies evaluated independently, by replication category. 28

2.4 Example: Confidence intervals for one original study and 6 replicates. 39

4.1 PRISMA flowchart of studies included in this chapter. 61

4.2 Amount ratio by Study Id for all question pairs. 64

4.3 Time delay ratio by Study Id for studies with a non-zero delay to the sooner amount. 66

4.4 Porportion of participants choosing the later amount across all studies. 67

4.5 Proportion choosing the later option for each study. 67

4.6 Proportion of participants choosing the later amount by amount ratio. 68

4.7 Proportion choosing later option by amount ratio coloured by sign for Faralla et al. (2012). 69

4.8 Proportion choosing later option by amount ratio coloured by sign for Hardisty et al. (2013) 70

4.9 Proportion choosing later option by amount ratio coloured by sign for Han and Takahashi (2012) 71

4.10 Proportion choosing later option by amount ratio coloured by sign for Xu et al. (2009) 72

4.11 Proportion choosing the later option by time delay difference. 73

4.12 Sign effect size for 832 question pairs across 9 studies. 76

4.13 Sign effect size for 832 question pairs by Study Id. 77

4.14 Density plot of amount ratio coloured by direction of sign effect size across all 9 studies. 77
4.15 Sign effect size by amount ratio for each study. 78
4.16 Sign effect size by amount ratio for Faralla et al. (2012) 79
4.17 Sign effect size by amount ratio for Hardisty et al. (2013) 79
4.18 Sign effect size by amount ratio for Han and Takahashi (2012) 80
4.19 Sign effect size by amount ratio for Xu et al. (2009) 81
4.20 Indifference point calculation illustration for Faralla et al. (2012): Choices for each participant on a single trial 83
4.21 Cumulative density of the number of switch points across all 30 trials for Faralla et al. (2012) 84
4.22 Illustration of a last minute switch. 84
4.23 Indifference point calculation illustration for Xu et al. (2009): Choices for each participant in a single trial 86
4.24 Cumulative density of the number of switch points for Xu et al. (2009). 87
4.25 Within-subject variation in discounting behaviour across all studies. . 91
4.26 Within-subject variation in discounting behaviour for each study. . . 92
4.27 Cumulative density of participant responses on question pairs, coloured by type of discounting behaviour 93
4.28 Sign effect size for each participant coloured by whether the participant had zero discounting for gains or zero discounting for losses. . . 95

5.1 The proportion of later choices for each participant, coloured by sign. 103
5.2 Density plots of the proportion of later choices coloured by sign for each participant. 104
5.3 The relationship between the proportion of later choices and the time interval split by sign and pattern of responses. 105
5.4 Boxplots of the proportion of later choices and gender by interval and sign. 105
5.5 The relationship between the proportion of later choices per participant and the sooner amount of money split by sign and pattern of responses. 107
5.6 The relationship between the proportion of later choices and sooner amount by sign, for each participant. 108

5.7 Proportion of later choices and amount ratio by sooner amount and sign. . . . 109
5.8 Proportion of later choices and amount ratio for each participant coloured by sign. . . . 110
5.9 Proportion of later gains and amount ratio for each participant by the different sooner amounts of money. . . . 111
5.10 Proportion of later losses and amount ratio for each participant by the different sooner amounts of money. . . . 111
5.11 Percentage of indifference points that could not be calculated for each participant. . . . 112
5.12 Cumulative density plots of the number of switching points for each unique combination of factors. . . . 113
5.13 Switching behaviour for each participant when the sooner amount was available today and the later amount in 14 days. . . . 114
5.14 Estimated random effects residuals for each participant. . . . 118
5.15 Predicted probabilities for model 1. . . . 121
5.16 Predicted probabilities for model 2. . . . 122
5.17 Predicted probabilities for model 5. . . . 122
5.18 Xu et al. 2009: Reproducing Figure 1 showing the proportion of sooner choices by sign. . . . 125
5.19 The proportion of later choices for each participant. . . . 126
5.20 The relationship between the proportion of later choices and the time delays for each participant. . . . 127
5.21 The relationship between the proportion choosing the later option and time interval split by sign and pattern of relationship. . . . 128
5.22 Boxplots of the proportion of later choices and gender by interval and sign. . . . 129
5.23 The proportion of later choices and amount ratio by sign 131
5.24 The relationship between the proportion of later choices and amount ratio for each participant. . . . 131
5.25 Percentage of indifference points that could not be calculated by sign for each participant. . . . 133
5.26 Cumulative density plots of the number of switching points in each unique factorial combination of sign and time delay. . . . 134

5.27 Switching behaviour for each participant when the sooner amount was available today and the later amount in 14 days. 135
5.28 Estimated random effects residuals for each participant. 136
5.29 Predicted probabilities for model 1. 138
5.30 Predicted probabilities for model 2. 139
5.31 Predicted probabilities for model 4. 139
5.32 Han and Takahashi 2012: The proportion of later choices for each participant. 142
5.33 The proportion of later choices by time delay. 142
5.34 Proportion of later choices for each participant and time delay coloured by sign. 143
5.35 The relationship between the proportion of later choices and amount ratio by gender. 144
5.36 The proportion of later choices for each participant and amount ratio coloured by sign. 145
5.37 Box plots of the indifference points by presentation order, faceted by sign and gender. 146
5.38 Box plots of the indifference points for each participant by sign. . . . 146
5.39 Boxplots of the indifference points for each participant by sign. 147
5.40 Estimated random effects residuals for each participant. 148
5.41 Predicted probabilities for null model. 151
5.42 Predicted probabilities for model 1. 151
5.43 Predicted probabilities for model 3. 152
5.44 The proportion of later choices for each participant. 155
5.45 Density plots of the proportion of later choices by sign and delay, coloured by gender. 156
5.46 The relationship between the proportion of later choices and amount ratio by gender. 157
5.47 Density plots of the ratio of the sooner amount to indifference points at the different delays, coloured by sign. 158
5.48 Estimated random effects residuals for each participant. 159
5.49 Predicted probabilities for model 1. 161
5.50 Predicted probabilities for model 2. 162

5.51 Predicted probabilities for model 4. 162
5.52 Hardisty et al. 2009 Experiment 1: The proportion of later choices for each participant. 166
5.53 Box plots of the proportion of later choices for gender by sign. 167
5.54 The relationship between the proportion of later choices and marital status by sign. 168
5.55 The relationship between the proportion of later choices and marital status by sign. 169
5.56 The relationship between the proportion of later choices and job type by sign. 170
5.57 The relationship between the proportion of later choices and age by gender and sign. 170
5.58 The relationship between the proportion of later choices and amount ratio by gender. 171
5.59 Each choice each participant made by sign 172
5.60 Density plots of the indifference points coloured by sign. 173
5.61 The difference in indifference points for gains and loss for each participant. 173
5.62 Estimated random effects residuals for each participant. 174
5.63 Predicted probabilities for null model. 177
5.64 Predicted probabilities for model 1. 177
5.65 Predicted probabilities for model 2. 178
5.66 Hardisty et al. 09 Experiment 2: Density plots of age for men and women. 180
5.67 The proportion of later choices for each participant. 181
5.68 Density plots of the proportion of later choices by sign, coloured by gender. 182
5.69 Proportion of later choices and age by sign. 183
5.70 The relationship between the proportion of later choices and amount ratio by gender. 184
5.71 Each choice each participant made by sign 185
5.72 Density plots of the indifference points coloured by sign. 186
5.73 Estimated random effects residuals for each participant. 187

5.74 Predicted probabilities for null model. 189
5.75 Predicted probabilities for model 1. 190
5.76 Predicted probabilities for model 2. 190
5.77 Odds ratio with confidence interval of the sign term for each study. . 193
5.78 Sensitivity and specificity with confidence interval from the ”best” model for each paper. 194

Acknowledgements

I would like to thank my three supervisors: Professor Daniel Read and Professor Jerker Denrell from the Behavioural Science Group, and Professor Jane Hutton from the Department of Statistics. Daniel and Jerker, for providing your subject-matter expertise. Jane, for patiently training me as a statistician and teaching me what selflessness, compassion, humility and courage mean through your actions.

I am grateful to have been part of the Statistics department for the last 4 years. I would like to thank the support and academic staff, especially Dr. Martine Barons, Dr. Linda Nichols, Dr. John Fenlon, Professor John Copas and Professor David Firth.

I would like to thank my fellow PhD mates who made me enjoy going into the office, especially the Mathematical Sciences Building level 4 tea time crew: James Griffin, Stefan Stein, Lewis Rendell and David Selby. You filled my days with laughter and joy. I would like to especially thank David Selby and Lewis Rendell for helping me to learn statistics, mathematics and programming over the 4 years.

For being great collaborators, I would like to thank Evan Fradkin, Anna Trendl, Anne Miloschewski, Dr. Lisheng He, Dr. Mark Fiecas and Professor Marcus Munafo.

I would like to thank my wonderful housemates, especially Marco Prantoni and Dr. Giacomo Zanella. It was always great to go back home knowing that I can speak with you about anything. Giacomo, I will always appreciate your evening lessons on probability!

I would like to thank Professor Nick Chater and Professor Petra Macaskill for igniting my interest in behavioural science and statistics. Your patience, encouragement, and passion for the subjects shaped my interests and inspired me to pursue this PhD.

I am grateful for funding and support from The University of Warwick Bridges–Leverhulme Doctoral Training Programme, which allowed me to pursue my interests in 'bridging' the behavioural science and statistics disciplines.

Finally, I would like to thank my family for their unwavering support and love. To my older siblings, Karin and Kelvin. And most importantly, to my parents. Thank you, Mum and Dad, for never giving up on me and for all the sacrifices you've made for me over my lifetime. This is my best attempt at being a doctor!

Declaration

This thesis is the result of my own work and research, except where otherwise indicated. This thesis has not been submitted for a degree at another university.

Abstract

Psychology and economics are undergoing a 'reproducibility crisis', with researchers attempting to replicate more published studies to verify existing findings. This thesis investigates the role of statistical practice in the reproducibility crisis. A 100-item checklist of recommended statistical practices was developed based on guidelines by the American Psychological Association. The checklist was used to evaluate a sample of psychology and economics studies that were already independently replicated. On average, the evaluated studies adhered to 30% of recommended statistical practice. Incomplete reporting hampered meaningful evaluation of the association between adherence to the checklist items and replication success.

Next, the thesis focusses on the sign effect, which is an established intertemporal choice anomaly. Verbal descriptions of the sign effect are formalised, a hypothesis testing framework is proposed and the concept of a discount rate is critically discussed. Then, the first systematic review and meta-analysis of the sign effect was attempted. Results suggested substantial heterogeneity within and between participants, which is not apparent when the convention of analysing aggregated data is used. There was a surprising amount of observations where no discounting occurred and where discount rates could not be estimated.

Then, individual participants' responses to questions are modelled to estimate the extent to which the outcome sign and other factors influence choices. Results suggested that the later amount was chosen more often for gains than losses. There was substantial heterogeneity within and between studies.

Good statistical practice is central to tackling the reproducibility crisis. Definitions need to be explicitly formalised, data need to be described sufficiently, assumptions need to be explored empirically, study designs need to be informative, and the different types of heterogeneity need to be documented and accounted for.

Chapter 1

Introduction

> *Do as I do when I reach for my glass of wine. Think: do I want the glass of wine or do I want to raise my own risk of breast cancer? I take a decision each time I have a glass.*— Professor Dame Sally Davies

Put yourself in the shoes of Professor Dame Sally Davies. There are costs and benefits that will occur at different points in time. Ask yourself: what *should* you do and what would you *actually* do?

This is type of decision, involving outcomes that occur at different points in time, is known as an 'intertemporal choice' in economics. How *should* you decide? According to traditional economic theory, by discounting future outcomes at a constant discount rate and then choosing the outcome with the highest value. Is that how people *actually* decide? Apparently not. Decades of experimental findings have documented numerous "anomalies", i.e. 'systematic deviations' from the traditional theory of constant discounting (Frederick, Loewenstein, and O'Donoghue 2002; Loewenstein and Thaler 1989).

The initial focus of this thesis was to conduct a meta-analysis of intertemporal choice anomalies. It is considered gold-standard practice in medicine to complement a meta-analysis with a systematic review, where the quality of included studies are appraised using a standardised checklist. In medicine, there is a collection of checklists, each

for specific purposes and study designs (Altman and Simera 2016). However, such standards and tradition have not found their way into the behavioural sciences yet.

Because no suitable behavioural science checklist was available, a checklist was developed in Chapter 2 and used to evaluate a sample of psychology and economics studies that were already independently replicated. The checklist was developed according to recommended practices from the American Psychological Association (Wilkinson and Task Force on Statistical Inference 1999), as well as neighbouring disciplines, such as animal studies, neuroscience and medicine. The checklist was developed under the constraints of a PhD thesis and may need to be published as a pilot study, with the view of being further developed by the larger community.

Chapter 3 introduces the sign effect anomaly in intertemporal choice. It attempts to formalise the definition of the sign effect with probabilities. It also critically discusses issues concerning, and assumptions of, the sign effect.

Chapter 4 provides a systematic review of the sign effect. This includes exploratory data analysis of aggregate data and individual participant data (IPD) from existing studies. The systematic review focussed on studies that present question pairs of binary choices involving monetary gains and losses to participants. A question pair is a pair of questions with the same amounts of money and time delays but differ in the sign (gain or loss). An example of a question pair is:

- Would you prefer to gain £100 today or gain £110 in a year?
- Would you prefer to lose £100 today or lose £110 in a year?

Chapter 5 models individual participants' responses to question pairs involving gains and losses. The chapter aims to investigate the factors that influence individual participants' choices in the presence of gains and losses. It also examines within- and between-participant heterogeneity. Finally, Chapter 6 concludes.

Chapter 2

Statistical Practice and Replication Success in Behavioural Science: An Evaluation of Published Studies

Abstract

Psychology and economics are undergoing a 'reproducibility crisis', with researchers attempting to replicate more published studies. Debate has focussed largely on specific issues like power, and p-values (e.g. the 2016 ASA's Statement). However, the full statistical process, including the design and conduct, which provides the context for analysis, has been surprisingly overlooked. Our team of behavioural scientists and statisticians developed a 100-item checklist of recommended statistical practice to evaluate a sample of 39 studies from a population of 100 psychology and 18 economics studies that were already independently replicated. On average, studies adhered to ~30% of recommended statistical practices: the median scores for the 30 psychology, and 9 economics studies were 26.3% (range: 14.5% to 37.8%), and 29.3% (range: 21.4% to 32.9%) respectively. Incomplete reporting hampered

meaningful evaluation of the association between the checklist scores and replication success. Ethical implications, and reproducibility concepts are discussed. Improving the quality of reporting in behavioural science will facilitate replication efforts. Our checklist provides a standardised template of essential information that should be reported. Such checklists are ubiquitous in, and have raised the standards of, medical research.

2.1 Introduction

> *In a sense, behavioural economics extends the paternalistically protected category of "idiots" to include most people, at predictable times.* — Camerer et al. (2003, 1218)

Governments around the world and international organisations, such as the OECD and World Bank, increasingly use findings from economics and psychology to inform policy decisions (Halpern 2015). Findings seem to consistently document human 'errors—apparent violations of rationality—that justify the need for paternalistic policies to help people make better decisions and come closer to behaving in their own best interest' (Camerer et al. 2003, 1218). However, recent evidence suggests that many published findings may not be replicable.

The Open Science Collaboration (2015) (OSC) and Camerer et al. (2016) attempted to replicate 100 'experimental and correlational studies' psychology studies, and 18 'between subject treatment comparison' economics studies respectively. These studies were sampled from three psychology journals (first 2008 issue) and two economics journals (up to August 2014) that are highly-influential. Only one key result within a study was chosen to be replicated as an article tends to have multiple studies and results. Only 35 out of 97 psychology studies, and 11 out of 18 economics studies were deemed to have 'replicated successfully', i.e. the replication result was in the original direction with $p < 0.05$.

Both the OSC and Camerer et al. (2016) attempted to explore multiple factors associated with replication success. These factors included characteristics of the

original and replication studies. The strongest correlation evidence was a rather weak association between replication success and the p-value of the original study, e.g. Spearman rank correlation coefficients of -0.33 (Open Science Collaboration 2015), and -0.57 (Camerer et al. 2016). However, the role of statistical standards underpinning the original studies was not examined.

The reproducibility debate has largely focussed on single issues, such as p-values (Benjamin et al. 2018; Wasserstein, Lazar, and others 2016), and statistical power (Anderson and Maxwell 2017; Ioannidis, Stanley, and Doucouliagos 2017) in individual studies. However, surprisingly little attention has been paid to the wider spectrum statistical issues (Loken and Gelman 2017). Statistical evidence encompasses a wide range of issues beyond p-values and statistical power.

Although there are no accepted guidelines to evaluate statistical standards in psychology and economics studies involving human participants, the established practice of using standardised checklists to evaluate medical research can be adapted. In medicine, clinicians and statisticians have developed standardised checklists to improve the statistical evidence of studies, and enable efficient replication efforts (Altman and Moher 2018; Begg et al. 1996). Having a standardised framework can contribute to a high-quality body of comparable evidence, which can then be used in meta-analyses. Thus, standardised checklists help facilitate a major goal of evidence-based medicine, which is to safely inform medical decisions using a collection of the highest-quality scientific evidence available.

It extremely timely to investigate the statistical standards underpinning published behavioural science studies, and its relationship with replication success. Behavioural scientists are increasingly collaborating, and expending significant resources, to replicate studies and better ascertain the veracity of published results (Camerer et al. 2018, 2016; Open Science Collaboration 2015). However, methodologically weak studies can produce biased estimates, and the results are not reproducible (Jüni, Altman, and Egger 2001). Researchers might reconsider dedicating limited resources to replicating a study if it were methodologically weak.

We aim to address two main questions with a descriptive and exploratory study.

To what extent does behavioural science research adhere to recommended statistical practices? What is the relationship between adherence to recommended statistical practices and replication success?

We make three major novel contributions. First, we developed a comprehensive checklist of recommended statistical practices for behavioural science studies involving human participants, which adapts guidelines, recommendations, and checklists from several disciplines, including medicine. Our standardised checklist also incorporates a report by Wilkinson and Task Force on Statistical Inference (1999), who were commissioned by the American Psychological Association (APA) to provide guidance on the use and reporting of statistical methods in psychology journals. Their guidelines are reflected in the APA's publication manual, which implies that the psychological community thinks these standards need to be followed.

Second, our team of behavioural scientists and statisticians used the checklist to evaluate the current state of statistical practice that underpin the original studies chosen for replication by the OSC and Camerer et al. (2016). Our checklist captures the wider spectrum of statistical practice, and broadly reflects the *Ethical Guidelines for Statistical Practice* (Committee on Professional Ethics of the American Statistical Association 2016). Third, we respond to calls by the OSC to investigate factors associated with replication success by exploring its association with adherence to recommended statistical practices in the original studies.

To summarise. Governments and international organisations increasingly apply behavioural science research, which posits that paternalistic policies are needed to correct irrational human decisions. However, the recent replication evidence from the two collaborative efforts in psychology and economics have called into question the reliability of established findings (Camerer et al. 2016; Open Science Collaboration 2015). In light of the increasing large-scale efforts to replicate studies, it is timely to understand the extent to which original studies adhere to recommended statistical practices, and its relationship with replication success. Our team of behavioural scientists and statisticians developed a novel checklist of recommended statistical practices, and used it to evaluate a sample of original studies the OSC and Camerer

et al. (2016) chose to replicate.

The rest of the chapter is structured as follows. Section 2.2 details the development of our checklist of recommended statistical practice for economics and psychology studies involving human participants. Section 2.3 describes our pilot study. Section 2.4 describes the methods of, and results from, our main study. Section 2.5 discusses the results, limitations, ethical implications, and fundamental reproducibility concepts across several disciplines, including a proposal for a new definition of reproducibility based on the concept of bio-equivalence in medicine. Section 2.6 provides direction for future research, and Section 2.7 concludes.

2.2 Developing a Checklist

A checklist of recommended statistical practice for economics and psychology studies involving human participants was initially developed in May 2016 and refined until November 2016. KTKL used the checklist to evaluate several studies, noted any associated difficulties, and made changes under the guidance of JLH, who also evaluated one study with the checklist.

The checklist had 100 items, which were predominantly adapted from the report by Wilkinson and Task Force on Statistical Inference (1999). The proposed guidelines were rephrased into questions suitable for a checklist. Table 2.1 provides an example of how the "Design" section in Wilkinson and Task Force on Statistical Inference (1999) was rephrased for inclusion in the "Design" section of the checklist. The items in the checklist were also informed by methodological debates, and other guidelines and checklists across multiple disciplines, such as animal studies (Kilkenny et al. 2009), economics (Hertwig and Ortmann 2001; Friedman and Sunder 1994), medicine (Higgins and Altman 2008; Downs and Black 1998; Begg et al. 1996; Altman et al. 1983; Gore, Jones, and Rytter 1977), neuroscience (Nichols et al. 2017), and psychology (American Psychological Association 2013).

Table 2.1: Example of how guidelines were rephrased for the checklist.

Wilkinson et al. (1999, p. 594): Design section	Checklist items: Design section
Make clear at the outset what type of study you are doing… There are many forms of empirical studies, including controlled experiments, observational studies, …	Study design specified by author(s).
Some are hypothesis generating… Some are hypothesis testing…	Was study hypothesis generating, testing, or both? Number of study hypotheses as specified by authors. Description of study hypotheses.
For studies that have multiple goals, be sure to prioritise those goals.	Were hypotheses, aims or goals clearly stated? If study had multiple hypotheses, aims or goals, were they prioritised?

A search was conducted to identify other relevant checklists by systematically searching through online databases and textbooks. Five electronic databases were searched from their inception to 05 October 2016 for relevant checklists or guidelines in economics or psychology.

The five e-databases were: *Scopus*, *Web of Science*, *PsycINFO*, *Econlit with full text*, and *ScienceDirect*. The exact search terms used can be found in the Appendix. The contents of six experimental economics and psychology textbooks were also briefly surveyed in October 2016 (Fréchette and Schotter 2015; Dunbar 2008; Cook, Campbell, and Shadish 2002; Kagel and Roth 1997; Friedman and Sunder 1994; Davis and Holt 1993).

No similar checklists for economics or psychology studies were found from the literature search. The literature search through five e-databases yielded 1,073 hits,

i.e. references. Duplicates were removed in EndNote (version X7). The titles and abstracts of the remaining 668 references were screened. This entire process was performed independently by KTKL and a psychology undergraduate student, and then cross-checked.

There were several recommendations and guidelines but no explicitly named checklists. The two most relevant articles found were Wilkinson and Task Force on Statistical Inference (1999), and the 'journal article reporting standards' in the publication manual of the American Psychological Association (2013). The six experimental economics and psychology textbooks were scanned and while there were methodological recommendations, there were no checklists.

Several related articles were also identified from KTKL's reading of the ongoing reproducibility debate (Wicherts et al. 2016; Brown et al. 2014; Simmons, Nelson, and Simonsohn 2011). Wicherts et al. (2016) provides the most relevant checklist with 34 items. It was published in November 2016, just after our systematic search. It could not have been identified in our search, and was not taken into account in this paper.

2.3 Pilot Evaluation of Statistical Practice

A pilot study was done from 08 November 2016 to 18 December 2016, for KTKL to gain familiarity with using the checklist, and refine the checklist items for the subsequent main study. A sample size of 30 psychology and 6 economics studies was determined in advance, based on the available resources. The target population was the 97 original psychology studies (Open Science Collaboration 2015), and 18 original economics studies (Camerer et al. 2016) chosen for replication. Figure 2.1 displays a diagram of the sampling frame.

The population was first divided into two groups: whether a study replicated "successfully" or not (as per the results reported by the OSC and Camerer et al. (2016)). Then, three strata were created: "Very successful", "Successful"" and "Unsuccessful" replication results. Stratified sampling was conducted with 10 psychology and

2 economics studies sampled in each of the three strata.

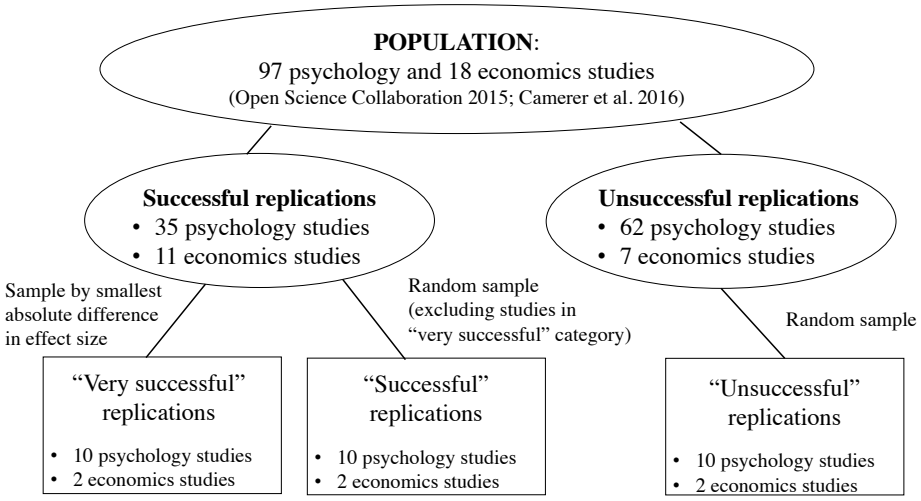


Figure 2.1: Sampling frame for a pilot evaluation of studies.

Studies in the "Very successful" replications stratum were successfully replicated, and had the smallest absolute difference in the original and replicated standardised effect sizes. Studies in the "Successful' replications stratum were drawn from a simple random sample, after excluding the"Very successful" studies from the population of successfully replicated studies. Studies in the "Unsuccessful" replications were also drawn from a simple random sample from the population of studies that did not replicate successfully.

An attempt was made to blind KTKL when evaluating the studies. The evaluation order was randomised by using the `sample (without replacement)` function in the statistical software, (version 3.4.3). The aim was to reduce the awareness of the strata to which each paper belonged. Results are summarised in Lim (2018).

2.4 Main study: Independent Evaluations of Statistical Practice

2.4.1 Methods

This study aimed to address two questions. To what extent does behavioural science research adhere to recommended statistical practices? What is the relationship between adherence to recommended statistical practices and replication success? We evaluated published studies using the checklist developed in Section 2.2, which also included an aspect on the time it took to evaluate a study. The studies were scored based on the extent to which they adhered to the checklist items.

The sampling frame was similar to that used in the pilot evaluation (see Figure 2.1), with the following two exceptions. First, the studies sampled and evaluated in the "Successful" and "Unsuccessful" replication strata were excluded from the population, and thus would not be evaluated here. The studies sampled in the 'Very successful' replication stratum were kept, and thus were evaluated here again. This meant that the population of "Successful" replications consisted of 25 (i.e. $35 - 10$) psychology studies and 9 economics studies. The population of "Unsuccessful" replications consisted of 52 (i.e. $62-10$) psychology studies and 5 (i.e. $7-2$) economics studies.

Second, we decided to increase the number of economics studies to be sampled from 6 in the pilot to 9, which represents 50% of the target economics population. Within each of the three strata, three economics studies were sampled. The number of psychology studies sampled remained the same at 30. There were 10 psychology studies and 3 economics studies from each of the three strata ("Very successful", "Successful", and "Unsuccessful" replications). This sample size of 39 studies was the maximum number of studies we could evaluate within the available resources and time frame.

Studies were first independently evaluated by two PhD students. KTKL evaluated all 39 studies. The 30 psychology studies were randomly assigned in a 1:1:1 ratio to

three other PhD students: EF, DAS, and AT, who each had 10 psychology studies to evaluate. These students have a background in economics, psychology, or statistics. The 9 economics studies were assigned to a PhD student with a background in economics and statistics: AM. Then, answers on the checklist between evaluators were cross-checked. Differences were discussed and resolved. An arbiter (MJAF) decided on any unresolved differences.

An attempt was made to blind all evaluators from the replication outcomes of the studies they were evaluating. The studies were evaluated in an order that was randomised. Evaluators were told not to check the replication outcomes from the data posted online by the OSC, and Camerer et al. (2016). The strata to which the studies belonged were recorded in a separate file, and only revealed after the evaluations were completed.

To gain familiarity with the checklist, all evaluators were given a few studies to evaluate initially. A meeting was held in March 2017 to resolve differences in understanding of checklist items. This resulted in changes to the checklist items, such as improving the clarity of certain items. The evaluations of the 39 studies continued until April 2018. Mid-way through the evaluations, four items were dropped because no common understanding could be reached. This is documented in Section 2.4.2.13 and Section 2.4.2.15.

To summarise the evaluation results from the many checklist items, a score was created. Each study was assigned a score based on $k = 55$ checklist items that had 'Yes' or 'No' categorical responses. The 55 items, and the remaining checklist items not included in the calculation of the score, are shown in the Appendix.

The score for the i^{th} study is calculated as

$$\text{Score}_i = \frac{\left(\sum_{j=1}^{k} \text{Y}_{i,j} + 0.5(\text{YS}_{i,j})\right)}{k - \sum_{j=1}^{k} \text{NA}_{i,j}} \times 100$$

where $\text{Y}_{i,j}$, $\text{YS}_{i,j}$, and $\text{NA}_{i,j}$ are indicator functions for the j^{th} checklist item having a response of 'yes', 'yes for some', and 'not applicable', respectively.

In the numerator, each 'yes' response earns 1 point while each 'yes for some' response earns half a point. All items were uniformly weighted because they facilitate interpretation, and are preferred for new instruments (Fletcher et al. 1992). In the denominator, the number of NA's is subtracted from the 55 checklist items included. For example, if a study had 20 'yes' responses, 10 'yes for some' responses, and 5 NAs, it would receive a score of 50. The score can range from 0 to 100. A higher score indicates greater adherence to recommended statistical practices.

The statistical software, (version 3.4.3), was used in the sampling, assignment of studies to evaluators, and subsequent analyses presented in this paper. For sampling, studies within each 'Successful', and 'Unsuccessful' replication category were assigned a number randomly drawn from a uniform distribution. Studies with the smallest 10 numbers were selected for evaluation. To randomly assign studies to evaluators, a random number was dawn from a uniform distribution and the `sample (without replacement)` function was used.

2.4.2 Results: Adherence to Recommended Statistical Practice

The median score for the 39 studies evaluated was 27.6 out of 100, with a range from 14.5 to 37.8. This implies that, on average, these studies were adhering to 28% of recommended statistical practice, as captured by our checklist items. The 30 psychology studies had a median score of 26.3 (range 14.5 to 37.8). The 9 economics studies had a slightly higher median score of 29.3 (range 21.4 to 32.9).

A strip chart of the scores by journal, with journals ordered by decreasing median values, shows the median scores across the three psychology journals were rather similar (Figure 2.2). *Journal of Experimental Psychology: Learning, Memory, and Cognition* had a median score of 27.6 (27.6 to 33.8), *Psychological Science* 26.3 (14.5 to 37.8), and *Journal of Personality and Social Psychology* 22.2 (15.5 to 37.5).

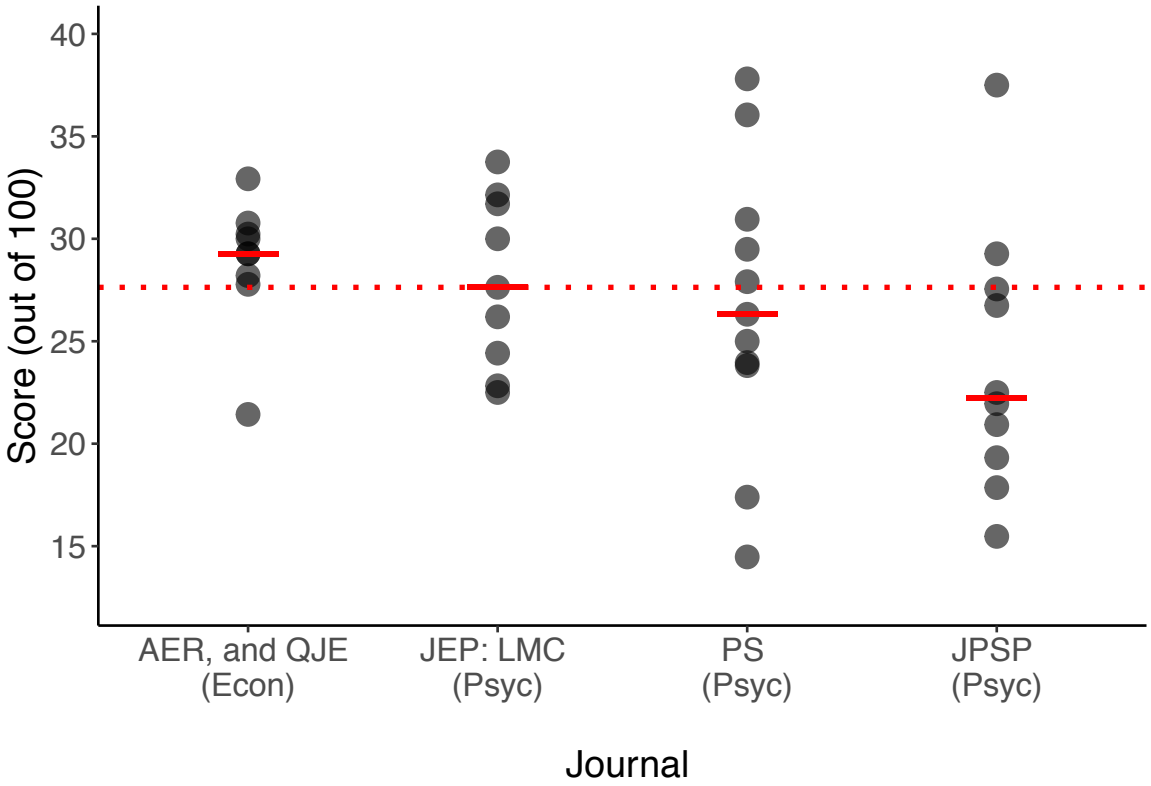


Figure 2.2: Strip chart: scores of 39 behavioural science studies evaluated independently, by journal. Each point is one observation. Journals are ordered by decreasing median score. Dotted horizontal line represents the overall median score for 39 studies. Solid horizontal lines represent the median score for each journal. The names of the journals are: American Economic Review (AER), Quarterly Journal of Economics (QJE), Journal of Experimental Psychology: Learning, Memory, and Cognition (JEP: LMC), Psychological Science (PS), and Journal of Personality and Social Psychology (JPSP). Two economics studies from QJE were combined with 7 economics studies from AER.

Table 2.2 presents the results of all 58 individual checklist items with 'Yes' or 'No' categorical responses. This includes 3 checklist items not included when calculating the score, which are marked with an asterisk. Items are grouped by the section of the checklist to which they belong, and are ordered by the number of 'Yes' responses recorded. Results for the "Very successful", "Successful", and "Unsuccessful" studies separately are in the Appendix.

Table 2.2: Table of 58 checklist items with 'Yes' or 'No' responses. Items are ordered by decreasing 'Yes' responses within each checklist section. A dash indicates that the categorical response was not applicable to that item. Three items with an asterisk were not included in the calculation of the score.

Question	Yes	Yes - some	No	Unclear	NA
Design					
Were hypotheses, aims or goals clearly stated?	37	0	2	-	-
If study had multiple hypotheses, aims or goals, were they prioritised?	6	0	5	10	18
Participants					
*Were participants recruited from a subject pool (potential to be recruited for multiple studies)?	17	0	5	17	-
*Were participants recruited solely for the purposes of this study?	5	0	4	30	-
Use your judgment: was sample representative of target population?	0	-	2	0	37
Measurement					
Did author(s) explicitly explain how each outcome variable was measured?	34	4	1	0	-
Did author(s) explicitly describe how each outcome variable relate to the goals of the study?	33	6	0	0	-
Did author(s) explicitly define each outcome variable?	33	5	0	1	-
If a physical apparatus was used, did author(s) describe the brand?	6	0	4	-	29
If a physical apparatus was used, did author(s) describe the model?	5	0	5	-	29

Table 2.2: Table of 58 checklist items with 'Yes' or 'No' responses. Items are ordered by decreasing 'Yes' responses within each checklist section. A dash indicates that the categorical response was not applicable to that item. Three items with an asterisk were not included in the calculation of the score. *(continued)*

Question	Yes	Yes - some	No	Unclear	NA
If a physical apparatus was used, did author(s) describe the design specifications?	3	0	6	-	30
If an instrument was used to collect data, did author(s) describe the validity with regard to the way the instrument is used in a population?	1	0	38	-	0
If instruments were used to collect data, did author(s) describe the reliability with regard to the way the instruments are used in a population?	0	1	38	-	0
Procedure					
Was sample size reported?	38	0	1	0	-
Did author(s) describe any anticipated sources of attrition due to noncompliance, dropout, death, or other factors?	2	0	34	3	-
Were pilot study sessions conducted?	1	0	1	37	-
Did author(s) describe personnel who collected the data?	1	0	15	-	23
Did author(s) describe personnel who administered the study?	1	0	35	-	3
Did author(s) report planning sample size in advance?	0	-	39	0	-
Did author(s) describe how such attrition may affect the generalisability of results?	0	-	39	0	-
If pilot sessions were conducted, was the purpose stated?	0	-	1	0	38

Table 2.2: Table of 58 checklist items with 'Yes' or 'No' responses. Items are ordered by decreasing 'Yes' responses within each checklist section. A dash indicates that the categorical response was not applicable to that item. Three items with an asterisk were not included in the calculation of the score. *(continued)*

Question	Yes	Yes - some	No	Unclear	NA
Allocation concealment					
Did author(s) explicitly specify randomising treatment to participants?	16	0	14	1	8
Use your judgment: Was allocation adequately concealed?	0	-	0	28	11
Deception					
Were participants informed of the true purpose of the study?	8	0	12	19	-
If deception was used, was it justified by authors?	1	0	11	-	27
Boredom					
Was there monetary incentive?	17	6	11	5	-
Blinding					
Were responses automatically captured, e.g. computer software programme?	25	1	5	8	-
Was an attempt made to blind participants to treatment group assigned?	3	0	1	21	14
If responses NOT automatically captured, was an attempt made to blind personnel collecting data?	0	-	0	14	25
Was an attempt made to blind personnel administering study?	0	-	0	32	7
Was an attempt made to blind those analysing the data?	0	-	1	38	-

Table 2.2: Table of 58 checklist items with 'Yes' or 'No' responses. Items are ordered by decreasing 'Yes' responses within each checklist section. A dash indicates that the categorical response was not applicable to that item. Three items with an asterisk were not included in the calculation of the score. *(continued)*

Question	Yes	Yes - some	No	Unclear	NA
Multiple testing					
Were methods used to handle multiple testing?	2	0	30	6	1
Attrition					
*Were there any missing data?	6	0	12	21	-
Did author(s) describe the participants excluded/dropped/lost to follow up	1	0	8	2	28
Were analyses compared with and without the participants excluded/dropped/lost to follow up?	1	0	8	2	28
Did author(s) describe differences with and without the participants excluded/dropped/lost to follow up?	0	-	8	2	29
Comparison group					
Was the number of of participants in each group stated?	29	0	8	2	0
Was sample size equally balanced across groups (+/- 10% difference)?	10	0	8	10	11
Did author(s) report the demography / characteristics of each group?	7	0	32	-	0
Were groups tested for baseline differences?	1	0	10	19	9
Analysis					

Table 2.2: Table of 58 checklist items with 'Yes' or 'No' responses. Items are ordered by decreasing 'Yes' responses within each checklist section. A dash indicates that the categorical response was not applicable to that item. Three items with an asterisk were not included in the calculation of the score. *(continued)*

Question	Yes	Yes - some	No	Unclear	NA
Was each statistical method described sufficiently to understand what was done? E.g. Spearman's rank correlation and not simply 'correlation'	11	14	14	-	-
Use your judgment: Was there evidence that data quality was checked (e.g. outliers, illegal values, "anomalies in the data")?	7	0	32	-	-
Use your judgment: Was basic data adequately described? (e.g. What was the distribution of the data? Was M +/- SD provided for normally distributed data? Was interquartile range or graphics provided for skewed data?)	1	0	32	6	-
Did author(s) justify the use of each statistical method?	0	10	29	-	-
Did author(s) describe the underlying assumptions of each analysis?	0	3	36	-	-
Use your judgment: do underlying assumptions seem reasonable given the data?	0	-	4	35	-
Did author(s) report examining residuals?	0	-	31	-	8
Were residuals presented graphically?	0	-	31	-	8
Results reporting					
Were interval estimates presented in each figure, where appropriate?	13	0	13	-	13
Were effect sizes presented for each finding?	10	22	7	-	0

Table 2.2: Table of 58 checklist items with 'Yes' or 'No' responses. Items are ordered by decreasing 'Yes' responses within each checklist section. A dash indicates that the categorical response was not applicable to that item. Three items with an asterisk were not included in the calculation of the score. *(continued)*

Question	Yes	Yes - some	No	Unclear	NA
Use your judgment: were all outcomes tested, reported?	9	0	9	21	-
Were actual p-values reported (e.g. 0.035 instead of p<0.05) for each finding except where the p-value is less than 0.001?	7	17	15	-	0
Is the data for the study available online?	6	0	17	16	-
Did the author(s) report the statistical software used?	4	0	35	-	0
Were interval estimates presented for each finding?	0	3	36	-	0
Was a protocol or pre-registered study mentioned?	0	-	39	-	0
Was there any deviation from the protocol/pre-registered study?	0	-	0	-	39
Discussion					
Were results generalised to the target population?	1	0	1	-	37

The results for the other items of the checklist not included in Table 2.2 are summarised below, by section of the checklist to which they belong. Most of these items had free text responses, and some had categorical responses that are different from those in Table 2.2.

2.4.2.1 'Design'

Aims, goals, or hypotheses. Almost all studies (37/39) had clearly stated aims. Eighteen had a single aim, goal, or hypothesis, and 21 studies had multiple aims, goals or hypotheses. Only 6 of these 21 studies prioritised their aims, goals or hypotheses.

Number and description of hypotheses. Half (22) the studies were judged to have specified one hypothesis to be tested or explored. Ten studies specified either two or three hypotheses, two studies specified either six or seven hypotheses, and two studies specified at least ten hypotheses. The remaining 3 studies were unclear. Evaluators commented that the number, and description, of hypotheses were not always reported explicitly. For a number of studies, there were discrepancies in what was recorded between the two evaluators.

Hypothesis testing or generating? Most (35) studies were judged to be 'hypothesis testing'; of these, six studies were 'both hypothesis testing and generating'. Three studies were 'hypothesis generating', and the purpose was unclear in the final study.

Study design as specified by authors. Generic designs were reported for half the studies, with 17 'Experiments', and 4 'Studies'. Of the remaining 18 studies, 8 studies were reported as 'Between-subjects', 3 studies as 'Within-subjects', 5 studies as 'Other', e.g. 'split plot factorial', and it was unclear in 2 studies.

2.4.2.2 'Participants'

Target population. Almost no studies stated their target population clearly. The samples in the only two studies that specified target populations were judged to not be representative of those target populations.

Inclusion/exclusion criteria. Only eight studies reported inclusion or exclusion criteria for participants to be eligible to participate.

Number, and description, of participants. Only 3 studies reported the number of participants invited to take part in the studies conducted, or the number of participants who agreed to take part in the studies they were recruited for. However, 38 studies reported the number of participants who took part in the studies conducted, and it was unclear in 1 study. Brief descriptions such as "undergraduates" were provided in 36 studies.

2.4.2.3 'Measurement'

Validity and reliability of instruments used. Although outcome measures were defined, only one study described the validity or reliability of the instruments with regard to the way the instruments are used in a population (Table 2.2). Evaluators have commented that studies usually provide some justification for the choice of instruments, which includes games, surveys, etc, although reliability and validity were not explicitly mentioned. Evaluators were also unclear about the meaning of the term, 'instrument'.

Estimated number of questions (rounds) to be presented (played) mentioned before results section of a paper. This was unclear in 1 study. In the remaining 38 studies, this ranged from 1 to 480 questions or rounds. The exact number of questions or rounds was not always explicitly reported in a study. Duration, the expected time to complete all questions, was not recorded.

2.4.2.4 'Procedure'

Sampling procedure. Although sample sizes were reported, information on attrition and personel was not (Table 2.2). Half (20) the studies provided some description of their sampling procedure, which includes descriptions such as, 'participants were recruited from the university's participant pool'.

Convenience or random sample? Convenience samples were used in 60% (24) of studies, one used a random sample, and the source of the sample was unclear in the remaining 14 studies.

Pilot study sessions. From Table 2.2, only 1 study reported conducting a pilot session before the main study, but it did not state the purpose of the pilot. It is worth noting that it is common, and even recommended, practice in behavioural science to conduct pilot sessions (Friedman and Sunder 1994).

Study setting. Most (35) studies reported the country in which their study was conducted, and the study setting, (31) e.g. laboratory, classroom, or online. It is worth noting that these were not explicitly reported in a number of studies, and evaluators had to infer this from information from other sections of the paper, e.g. the authors' affiliations for country.

Study start and end dates. Only 6 studies reported the start and end dates.

2.4.2.5 'Allocation and concealment'

Was treatment explicitly randomised? This question was not applicable to 8 studies (Table 2.2). Of the remaining 31 studies, half (16) explicitly reported randomising treatments. This item required evaluators to determine the treatment, and whether it was explicitly specified in a study.

Random assignment sequence generation procedure. Of the 31 studies for which this question was applicable, the procedure to generate the random assignment sequence was not reported or was unclear in 26 studies. The remaining 5 studies reported the following procedures: based on the order in which participants arrived at the lab, an odd and even number ordering of participants, flipping a coin, and using a computer.

2.4.2.6 'Deception'

Evidence of deception. Participants in only 8 studies were informed of the study's "true" purpose. Some examples that were used during the evaluation as evidence

of deception, which are slightly altered to preserve anonymity, are: 'participants answered a questionnaire that was intended to render credibility to the cover story but was not analysed'; 'the ostensible survey'; 'there was a surprise test'; and 'participants were not informed of the impending test'.

2.4.2.7 'Boredom'

Estimated number of questions (rounds) to be presented (played) mentioned in the results section of a paper. The number of questions or rounds ranged from 1 to 480 (85% of the 39 studies were estimated to have presented 10 or more questions/rounds). In many situations, this information was inferred or based on what was presented in earlier sections.

Monetary incentive. Nearly half (17) studies gave all participants monetary incentives; a further 6 gave monetary incentives to some participants.

2.4.2.8 'Blinding'

Two-thirds (25) of studies recorded participants' responses automatically, but there was almost no use of blinding (Table 2.2).

2.4.2.9 'Multiple testing'

Total number of significance test reported. Only two papers adjusted for multiple testing, although the median estimated number of significance tests reported was 12 after excluding 1 "Unclear" study (Min. 1, 1st Qu. 7, 3rd Qu. 30, Max. 300). The number of significance tests was not always clearly reported, so had to be estimated. Significance tests of coefficient estimates in regression models are included.

2.4.2.10 'Attrition'

Evaluators faced difficulties in accurately recording checklist items in this section. The checklist items in this section were designed for a single analysis. However,

studies usually conducted multiple analyses. Further, while studies may not have dropped entire participants from the analysis, it was not uncommon for studies to drop parts of the participants' responses. Studies handled their *missing data* by deleting or excluding the observations. This was difficult to capture in the checklist. Six studies reported some data as missing (Table 2.2), but only one gave any detail.

Number of participants included in analysis. The number of participants included in the analysis was unclear in 3 studies. In the remaining 36 studies, the number of participants ranged ranged from 4 to over 220,000. The number of participants included in the analysis was not always explicitly stated and inference had to be made, e.g. by using the degrees of freedom reported. Further, this number could vary within a study as it was common to report multiple statistical tests, e.g. we have 'between 23 and 374 participants' recorded.

Number of participants excluded from analysis. Half (19) the studies did not exclude any participants in the analysis. For two, no participants were excluded but data, e.g. responses, were dropped. It was unclear, or not reported in 9 studies if any participants were excluded, dropped or lost to follow up. In the remaining 9 studies, varying numbers of participants were excluded, dropped or lost to follow up.

Reasons for attrition given by authors. Some examples include: participants making 'either no incorrect or no correct responses for a particular item'; participants 'not following instructions'; participants having a reaction time outside a certain range; participants not responding to surveys or questionnaires; and participants responding with values of 0, which get dropped when transformed on the logarithmic scale.

2.4.2.11 'Risk of bias: Other'

Potential sources of risk of bias. Examples encountered while evaluating studies include: changing degrees of freedom across statistical tests without explanation, post-hoc subgroup analyses, a χ^2 test with only one observation in a cell, restricting analyses to only correct or certain responses, using different types of statistical tests without explanation (e.g. t-, F-, and χ^2 tests), interviewer-response bias (e.g. 'participants were tested individually… asked to report orally to the experimenter, on

each target occurrence'), and conflicts of interests (e.g. having a sample size of 4 participants, of whom 2 were the study authors).

Another issue worth mentioning is the generalising of findings from surrogate outcomes in a laboratory to real world outcomes. For example, Kessler and Roth (2012) looked at organ donation and allocation policies. The abstract stated that this would be modelled through an 'experimental game' in the laboratory, and that based on the results found, a certain allocation policy 'has a significant positive impact on registration'. The Experimental Design section explains a game where participants were informed in 'abstract terms' that they were trading "A" and "B" units. The Discussion section cautions against extrapolating results to 'complex environments outside the lab' in contrast to the statement quoted above from the abstract. This is further elaborated in the Discussion section.

Note, these are only referred to as *possible* sources of risk of bias, and should be treated as such.

2.4.2.12 'Comparison group'

Total number of groups. A third (13) of studies had one group, and another third (12) had two groups. Eight had 3 or 4 groups, six 6 or 7 groups and two had at least 10 groups. The number of participants in each group was reported by 29 studies, but only seven studies summarised group demography.

2.4.2.13 'Analysis'

Few (7) studies provided evidence that data quality was checked, and only one described the data adequately. No study were judged to have justified the use of each statistical method, described the underlying statistical assumptions, used reasonable statistical assumptions for the data, or checked residuals where applicable.

The following 3 checklist items were dropped mid-way through evaluations due to differences in interpreting the question. *What was the unit used in the statistical analysis? Were there repeated measures? Was a repeated measures analysis used?*

2.4.2.14 'Results reporting'

No study mentioned a protocol or pre-registered plan. Most (32) reported at least some effect sizes and actual p-values. However, interval estimates were rarely given, and only a quarter (9) of studies were judged to have reported all outcomes tested.

Interval estimates in figures. A common type of figure encountered was the plunger plot, even though it goes against recommendations of displaying the distribution of the data by (Wilkinson and Task Force on Statistical Inference 1999).

2.4.2.15 'Discussion' section of checklist

Were results generalised to target population? Only 2 studies that specified a target population, and one generalised results to the target population.

One checklist item was dropped mid-way through evaluations due to differences in interpreting the question: *Was a generic generalisation statement made?* However, it was common to encounter the use of vague language to describe participants in many discussion sections, e.g. "people" or "agents", without appropriate disclaimers, which seemed to imply that results applied beyond the immediate (convenience) sample.

2.4.3 Results: Adherence to Statistical Practice and Replication Success

There was little difference in median scores across the three replication categories, with studies that replicated "successfully" scoring slightly higher than studies that did not (Figure 2.3). The median score was highest in the "Successful" category: 29.3 out of 100 (range: 17.4 to 33.8), and 28.2 (range: 21.4 to 37.8) in the "Very successful" category. In the "Unsuccessful" category, the median score was 26.7 (range: 14.5 to 36). From the Appendix Table, the proportion of 'yes' responses was decreasing from "Very successful" to "Unsuccessful" categories in 17 of 58 questions, but only strictly decreasing for three.

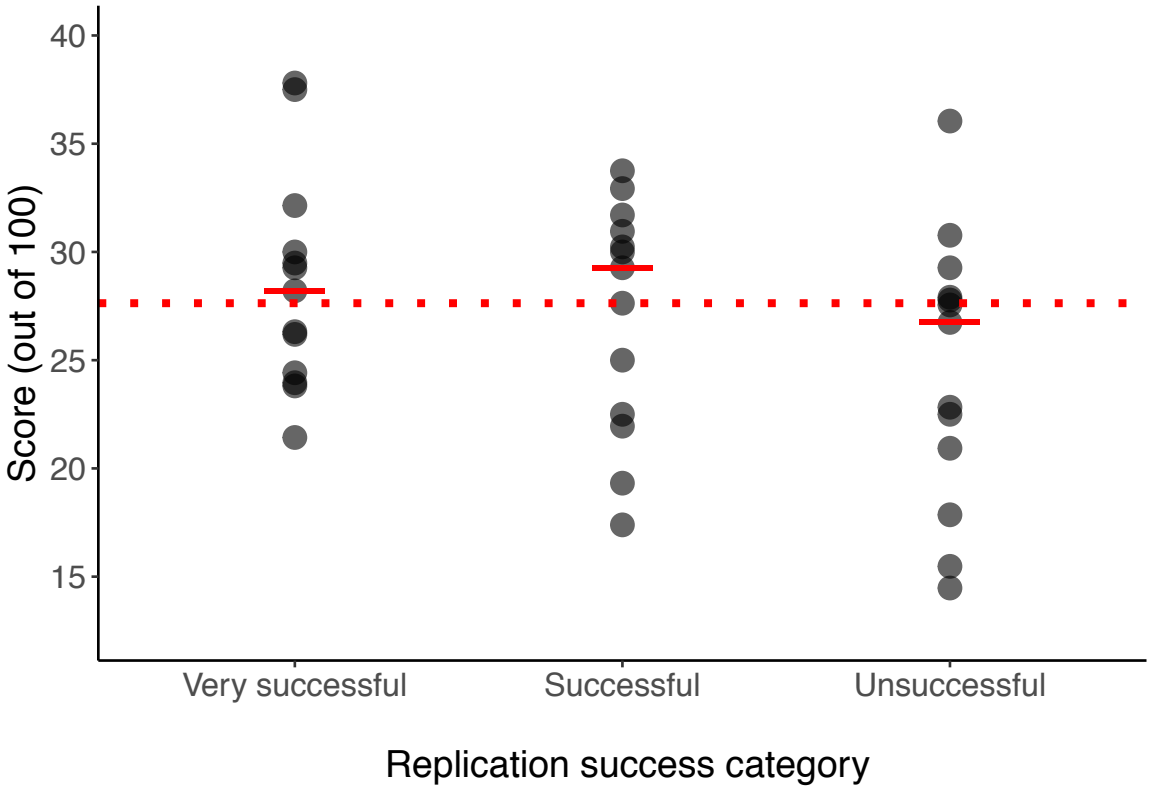


Figure 2.3: Strip chart: scores of 39 behavioural science studies evaluated independently, by replication category. Each point is one observation. Dotted horizontal line represents the overall median score for 39 studies. Solid horizontal lines represent the median score for each replication category.

2.4.4 Results: Evaluation Time

The time taken to evaluate a study was calculated by taking the mean time taken for both evaluators. The time for 31 studies will be reported in this section as one evaluator did not record the time taken for 8 out of 10 psychology studies.

The median time to evaluate 21 psychology studies was 49.0 minutes (range: 30.5 to 79.5 minutes, with 8 missing data points). On average, KTKL and the other evaluators took similar time for psychology studies, with median difference 8.0 minutes (range: 3 to 58.0 minutes). Economics studies took about an hour to evaluate: the median was 63.5 minutes (range: 48.0 to 151.0 minutes). KTKL took more than half an hour longer to evaluate an economics study than AM: median 38.0 minutes

(range: 7 to 69.0 minutes).

2.5 Discussion

The median score was 27.6 out of 100 for the 39 original studies (range: 14.5 to 37.8), implying that, on average, the evaluated studies adhered to 28% of the recommended statistical practices in our checklist, and no study adhered to more than 38% of the recommended statistical practices. The median score was 26.3 out of 100 (range: 14.5 to 37.8) for the 30 psychology studies, and 29.3 out of 100 (range: 21.4 to 32.9) for the 9 economics studies, suggesting little difference between the two disciplines. The median scores for the evaluated studies that replicated with "Very successful", "Successful" and "Unsuccessful" results were 28.2, 29.3 and 26.7 (out of 100) respectively, suggesting no apparent relationship between the checklist scores and replication success.

Overall, there was a lack of essential information reported across all the studies. For example, studies tend to report generic designs: 20 studies were reported as 'Experiment' or 'Study', and it was unclear in 2 studies. In 31 applicable studies, only 16 studies reported explicitly randomising treatments (conditions) to participants. These results suggest explicit and standardised guidance is needed on reporting essential study design information. Informative reporting of designs can facilitate the development of specific checklists in the future.

Accurate, transparent, and complete reporting of studies is crucial as published articles are a primary source of information and knowledge for the research community. If insufficient information is reported, researchers cannot critically evaluate the scientific methods and merits of a study. The utility of studies for the purposes of research synthesis, meta-analysis, and understanding how study characteristics are associated with replication success is diminished. It would also not be possible to replicate studies, and assess whether results are reproducible: the OSC and Camerer et al. (2016) coordinated closely with the original study authors, expending significant resources in the process, which is an inefficient approach in the long run.

The comprehensive checklist our team of behavioural scientists and statisticians developed provides a standardised template for what essential information should be reported in studies involving human participants. We extend other smaller checklists in behavioural science (Wicherts et al. 2016; Brown et al. 2014; Simmons, Nelson, and Simonsohn 2011). Our checklist adapts recommended practices, guidelines, and checklists from neighbouring disciplines, while incorporating many aspects from the report by the American Psychological Association's (APA) Task Force on Statistical Inference (Wilkinson and Task Force on Statistical Inference 1999). Such checklists are widely used in, and have raised the standards of, medical research (Turner et al. 2012).

Our practical checklist can benefit various stakeholders in the behavioural science community, and responds to calls for 'simple evidence-based editorial policies' that 'can improve reproducibility of science' (Camerer et al. 2016, 1436), and solutions to 'help researchers and readers to understand and communicate evidence more accurately' (Benjamin et al. 2018). Journal editors and reviewers can efficiently ascertain the extent to which the manuscripts they receive report essential information, which is an accepted practice in medicine. Funding agencies can evaluate proposals prospectively, and published studies retrospectively. Researchers can evaluate studies and decide which are worth replicating or building on, or use the checklist as a guide to decide which essential information needs to be considered and reported during the planning, pre-registration, and manuscript write-up stages.

Our results—from the first evaluation of statistical practice with a comprehensive checklist in the sample of studies chosen for replication by the OSC and Camerer et al. (2016)—contribute to ongoing calls in behavioural science to improve the quality of reporting (Christensen and Miguel, n.d.; Munafò et al. 2017; Camerer et al. 2016; Spellman 2015; Simmons, Nelson, and Simonsohn 2011). Systematic areas of strength and weakness in reporting and statistical practice standards across psychology and economics were documented in fine detail in the Results section. The community can use our results to have an informed debate about specific areas for improvement, and devise tailored solutions.

Our results broaden the debate that has largely focussed on p-values, power, and replication efforts (Camerer et al. 2018, 2016; Benjamin et al. 2018; Anderson and Maxwell 2017; Ioannidis, Stanley, and Doucouliagos 2017; Wasserstein, Lazar, and others 2016; Open Science Collaboration 2015). These are important issues to consider but they only form a part of the wider spectrum of statistical practice, which includes but is not limited to: question formulation, design, measurement, conduct, analysis and interpretation (Cox and Donnelly 2011). Our checklist captures aspects from this wider spectrum of statistical practice, and our results provide a more holistic picture of the state of statistical practice in economics and psychology. We hope this shifts the debate to consider a wider range of statistical issues.

Finally, we extended previous work exploring the association between characteristics of the original studies and replication results (Camerer et al. 2016; Open Science Collaboration 2015). Our results did not suggest any association between the checklist scores and replication success. However, this could be due to the incomplete reporting across all studies, and lack of variability in the scores: no study scored above 40 out of 100. The lack of essential information reported will hinder other attempts to explore how study characteristics relates to replication success.

Our study had the following limitations. First, an instruction manual for using the checklist to evaluate studies has not yet been developed, which contributed to inconsistent evaluations. Although our checklist adapted guidelines and checklists from several disciplines, information from other disciplines may be context-dependent, and guidelines may be insufficient for the purposes of using in a checklist. Having multiple evaluators meant variation in interpreting certain, more subjective, items in the checklist. In fact, 4 items were dropped midway because evaluators could not reach a consensus.

Second, the checklist needs further refinement. A checklist with 100 items could introduce recording errors due to boredom and genuine mistakes. In medicine, checklists tend to be shorter and for specific study designs, e.g. the CONSORT statement for randomised controlled trials (Begg et al. 1996). This would make the checklist more informative, e.g. by having more applicable questions, and practical to use

on a large number of studies. However, it is important to note that the novelty of this current checklist also provided clear evidence that, unlike medicine, it was common for study designs to be generic or not explicitly reported, e.g. 'experiment'. As such, researchers might run into the issue of not knowing which specific study design checklist to use.

A checklist can only assess what was reported, which may not reflect what was actually done although a reasonable view to hold is, 'if it is not reported, it was not done'. Reporting of information also depends on page limits, reporting norms, and assumed audience knowledge of different journals. However, reporting standards proposed by the APA were meant to encourage consistent reporting across psychological disciplines, and 'promote interdisciplinary dialogue' (APA Publications and Communications Board Working Group on Journal Article Reporting Standards 2008, 846).

The third limitation of this study related to our sample. Our population of interest was limited as replications are not commonly published (Makel, Plucker, and Hegarty 2012; Smith 1994). Our sample size of 39 studies was based on available resources. Our sample was limited to a subset of studies published in a few influential journals in economics and psychology. However, the journals in this statistical population are likely to greatly influence the behavioural science literature, and policy decisions. Based on our sampling frame of 30% of the psychology population (97 studies) and 50% of the economics population (18 studies), we believe that the median score of 27.6 out of 100, and a maximum score of less than 40, are representative of our population.

Fourth, we assumed that the replication results were 'the truth'. Although the OSC and Camerer et al. (2016) coordinated closely with study authors and pre-registered the replications, the replicated studies could also be biased in the same way as the original studies. Finally, the OSC (2015) and (Camerer et al. 2016) chose a single key result from each study to replicate, but we evaluated the entire study, which typically reported more than one result, as it was seldom possible to only evaluate specific aspects related to a single result.

Finally, the association between responses to individual questions and replication

success was not formally examined. This was assessed visually and there were no clear patterns from inspecting the responses on each question by the three different replication success categories. Colleagues have suggested using shrinkage methods, such as the Lasso, to formally examine this relationship, which was not within the scope of this chapter.

2.5.1 Standards, Current Practices, and Ethical Implications

In response to our results, the behavioural science community might wish to consider the following questions relating to their standards, practices and ethics.

What does the behavioural science community consider to be acceptable standards of statistical practice? Our results demonstrate a considerable gap between expected standards that have been in place since the late 1990s, and actual practice. Many of our checklist items were adapted from the guidelines published by Wilkinson and Task Force on Statistical Inference (1999), who were commissioned by the American Psychological Association (APA). As these standards are reflected in the APA Publication Manual (2013), and explicitly encouraged for adoption (APA Publications and Communications Board Working Group on Journal Article Reporting Standards 2008), the community thinks these standards need to be followed.

Could behavioural science benefit from a greater involvement from the statistics community? Economists have suggested that their methodological practices are better than those found in psychology, and should be followed (Camerer et al. 2016; Hertwig and Ortmann 2001). Our results suggest there is negligible differences in statistical practice standards between economics and psychology. Both disciplines could look to evidence-based medicine, where a core feature is a close collaboration between medical statisticians and clinicians to answer clinically relevant questions. However, statisticians interested in social science must collaborate with, and embed themselves in, the community to the point of considering themselves to be social scientists (Goldstein 1984).

What are the ethical implications of using studies that do not adhere to recommended statistical guidelines to influence policy decisions? In medicine, it is considered unethical to use studies of poor statistical quality to inform medical practice (Hutton 1995). Is it still acceptable to claim, and operate on the belief, that behavioural science research findings 'justify the need for paternalistic policies to help people make better decisions and come closer to behaving in their own best interest' (Camerer et al. 2003, 1218)?

The incomplete reporting, and failure to adhere to recommended statistical practices, imply that current research practices in behavioural science are unethical. It is difficult, if not impossible to critically appraise if studies have yielded 'fruitful research for the good of society' (British Medical Journal 1996), or are of 'sufficiently high quality and robustness' (British Psychological Society 2010). This 'can lead to misleading information being promulgated and can have the potential to cause harm' (British Psychological Society 2010). Uncritically implementing unverifiable research findings in the real world is unwise, and unethical if more harm were done than good.

The *Ethical Guidelines for Statistical Practice* (Committee on Professional Ethics of the American Statistical Association 2016) starts by clarifying that 'because society depends on informed judgments supported by statistical methods, all practitioners of statistics… have an obligation to act in good faith, to act in a manner that is consistent with these Guidelines, and to encourage others to do the same'. 'Good statistical practice is fundamentally based on transparent assumptions, reproducible results, and valid interpretations'. We highlight several relevant recommendations, which reflect the broader ethical issues. An ethical statistician:

- 'Acknowledges statistical and substantive assumptions made in the execution and interpretation of any analysis'
- '… conveys the findings in ways that are both honest and meaningful to the user/reader'
- 'When reporting analyses of volunteer data or other data that may not be

representative of a defined population, includes appropriate disclaimers…'

- 'Strives to avoid the use of excessive or inadequate number of research subjects… by making informed recommendations for study size'
- 'Ensures that all discussion and reporting of statistical design and analysis is consistent with these Guidelines'
- 'Strives to promote transparency in design, execution and reporting or presenting of all analyses'

2.5.2 Replication, Repeatability, and Reproducibility

Without sufficient information being reported, it is not possible to replicate studies. If studies cannot be replicated, it is not possible to assess whether results are reproducible. We use "replication" to refer to repeating an observational or experiment study using the same inclusion and exclusion criteria, interventions, measurements, and analyses. That is, "replication" refers to the collection and analysis of data, in contrast to "reproducibility" of results. "Reproducible" results are results obtained from a replicated study, which are consistent with previously reported results within a reasonable level of uncertainty.

Definitions of "replication", "repeatability" and "reproducibility" differ between disciplines (Goodman, Fanelli, and Ioannidis 2016), perhaps more for pragmatic reasons than philosophical disagreements. Some agreement, particularly on the interpretation of different sources of variation (or "error"), would aid inter-disciplinary cooperation.

In behavioural science, the OSC uses these terms differently, perhaps inconsistently. 'Replication' refers to evidence: 'Scientific claims should not gain credence because of the status or authority of their originator but by the replicability of their supporting evidence' (p. aac4716-1). Instead of recognising variation as an intrinsic aspect of reality, and taking it into account in defining "reproducibility", variation is used

to excuse "irreproduciblity": 'Even research of exemplary quality may have irreproducible empirical findings because of random or systematic error' (p. aac4716-1). Inherent variation is then recognised by considering consistency, with "reproducibility" describing the extent to which consistent results are observed when scientific studies are repeated within scientific fields such as cancer studies. The authors' intention is to compare the overall replication rate with the average statistical power across studies if power were reported, although the chosen definition of "consistency" was undefined. For individual studies, several criteria are used to define successful replication: p-values, effects sizes, subjective assessments of replication teams, and combining effect sizes from the original and replicated studies.

Some concepts from other disciplines which might advance discussion in social sciences are given here. Bio-equivalence, a concept used in the regulation of medicines and pharmacokinetics, might provide a useful approach to thinking about reproducibility of an effect size in the social sciences. It allows for the reality of variation between people in assessing whether different preparations of a drug are effectively equivalent, or a new drug is sufficiently similar to an existing drug. Regulatory definitions typically require 90% confidence intervals for the ratio of effects to lie within an acceptance range of 80% to 125%. This treats the two drugs equally, but does assume there are robust methods for estimating the distribution of the ratio (Flouri et al. 2017), and that the experimenters are able to give both drugs to a suitable study population.

In engineering and manufacturing, repeatability and reproducibility are regarded as the two components of precision in a measurement system assessed by Gauge R & R analysis. Variation in a measurement system is attributed either to the device (instrument) used for measurement or to the people taking the measurement. In this context, "repeatability" is the variation in measurements when one person uses the same instrument to measure the same object several times. Variation arising from differences in the way people use the same instrument is termed "reproducibility". All manufacturing companies are expected to provide this information for audits, and the objects tested should be representative of the production processes. The adequacy of the measurement system is judged by the variation between measurements, by

instrument and person.

The concepts extend beyond standard Gauge R & R to comparing the performance of sets of instruments over time and in different laboratories. Preparation, storage and transport of biological samples can also give rise to variation. Regulatory authorities and academic researchers assess inter-laboratory variation, as differences can have serious consequences. Obviously, convictions or acquittals for offences such as driving while drunk should use reliable standards, and should not depend on the laboratory which carried out the analyses. Design of experiments and analysis of variance are long-established statistical techniques, first used to explore the factors which affect crop yields, but essential in understanding variation in measurement (Ellison, Barwick, and Farrant 2009). Analysis of variance is not merely a process for generating p-values.

Of course, social science has to comprehend many more sources of variation than engineering and physical sciences, and rarely has measurements as simple as the length of a bolt. The theories of measurement which underlie physical and social science are different (Hand 2004).

We suggest a definition of reproducibility of an effect in social science which is informed by bio-equivalence, but does not require repeated measurements on the same individual or units. The main idea for bio-equivalence is that there is no more than 10% probability that either effect size is 25% larger than the other. We illustrate the definition before formalising it and considering extensions.

The following assumptions are required. The original data, X_o, is a random sample (simple or otherwise) of a specified target population. A replication study should be from the same target population. The original study is assumed to have a target population, though this might only be implicit from the discussion about the extent to which the reported findings can be generalised. The target population is that for which the answer to particular questions is required, or estimates of particular effects, so that our inference can validly be generalised to that population. Under a common assumption in behavioural science that there are no differences between people (Henrich, Heine, and Norenzayan 2010b), the target population includes ev-

eryone. A replication study should have at least the same power as the original study; having the same number of units is a simple approximation for the same power. We also assume the replication study units are independent of the original units.

Let $\hat{\mu}_o = T(X_o)$ be the estimated effect in the original study, where $T(X_o)$ is the estimator based on the study data recorded for n people, X_o, and let the estimated variance of $T(X_o)$ be $\hat{\tau}_o^2$. The parameter of interest, μ, is well defined, and the statistic $T(X_o)$ is an unbiased estimator of μ. The population variance is σ^2. Alternative requirements for the estimator are that $T(X_o)$ is asymptotically unbiased and consistent. Without loss of generality, we assume the parameters of interest are means, though regression coefficients are also common. The variance of $T(X_o)$ is then $\tau_o^2 = \sigma^2/n$. Given data X_r and $\hat{\mu}_r$ from a replication study of size m, we consider the overlap of the confidence intervals (CI) (90% confidence intervals are used in bio-equivalence). If the replication CI is within that of the original study, and includes $\hat{\mu}_o$, the effect size has been reproduced: see Figure 2.4, original study and replicate 1. If there is no overlap (replicate 2), the study has not been reproduced. These are the two obvious cases.

From Figure 2.4, a replicate (sample number 3) with CI contained by the original CI, but which excludes the original $\hat{\mu}_o$, still reproduces the original result successfully. A very wide replicate CI (replicate 4) that overlaps the original CI would generally not be convincing, which is why the size or power of a replication study is important ($m \geq n$). Replicates 5 and 6 overlap the original CI but neither interval completely contains the other. A criterion for sufficient overlap is necessary. As replicates 5 and 6 are symmetric about $\hat{\mu}_o$, the criterion should give the same conclusion for both. The obvious approach is to consider the difference in means, $D(X_o, X_r) = \hat{\mu}_o - \hat{\mu}_r$, which has mean 0 and variance $\sigma^2/n + \sigma^2/m$ under the assumptions above. Reproducibility implies that the confidence interval for the difference must include zero.

As with engineering tolerance limits, the magnitude of permissible variation will depend on the context. The limits of the confidence interval for a ball-bearing might be in microns, whereas for the length of a plank of wood, millimetres would suffice.

When estimating the mean age at which a child can read a given story, one might wish the limits of the confidence interval of the difference in means between original and replicate studies to be within 1 month.

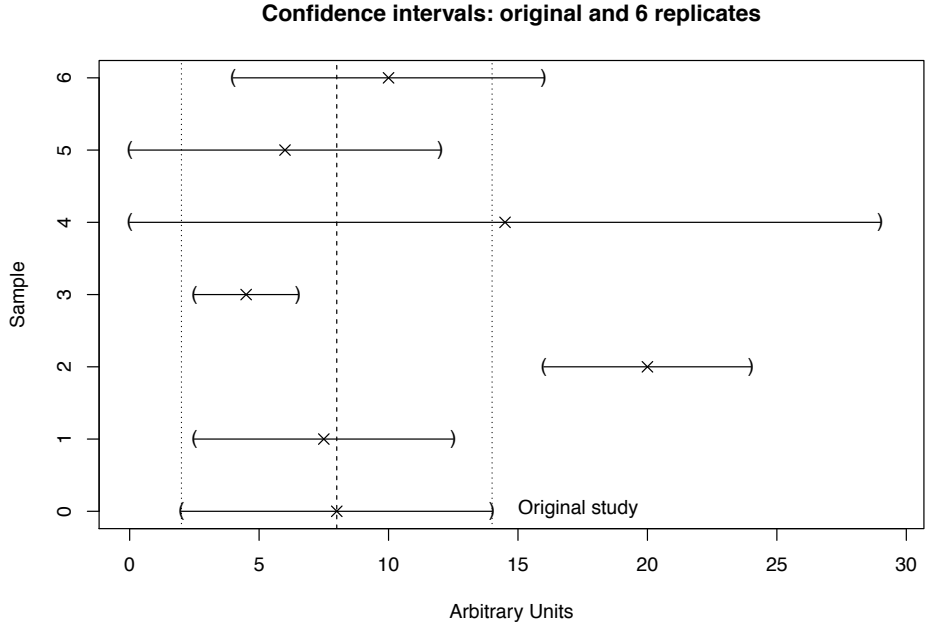


Figure 2.4: Example: Confidence intervals for one original study and 6 replicates.

Context is also important in deciding what the relevant statistics are and which statistics support the study conclusions. Suppose the original study estimated the age at which English mother-tongue and Chinese mother-tongue children can read a particular story, and concluded that girls had a lower mean age than boys. We might wish a replicate study to estimate the mean age for children, or for girls and boys separately. In the latter case, we should consider both the estimated mean for girls and boys being close to the original estimates, as well as the estimated difference between means for girls and boys being close to the original difference. This can be extended to language differences. That is, we are using analysis of variance.

A further consideration is how closely one expects to replicate an indirect conclusion. It is important to consider whether what is measured is of direct interest or is intended to capture an underlying quantity, such as generosity, or willingness to

donate. In drug development, a surrogate endpoint is a measurement used instead of an outcome that is important to a patient. Surrogate endpoints are used when the desired measurements are expensive or difficult to obtain, or take a long time to study. If tumours in people treated with a new drug shrink, it might be assumed the patients will live longer. Some surrogate endpoints have been validated by the Food and Drug Administration[1].

There should be a strong association between measurement and outcome. For example, if a new drug shrunk kidney tumours by half within a fortnight in mice, and the study concluded that it would increase time from diagnosis of kidney cancer to death in people, readers might be sceptical. Attempting to replicate the effect on kidney tumours in mice, to reproduce the estimated effect of halving tumour volume within a fortnight should be straightforward. Conclusions about increased life expectancy in kidney cancer patients could not be reproduced, as estimates of life expectancy are not part of the original study. Even measurements which are highly correlated with the desired outcome will not always provide reliable information.

We explore this approach to a definition of reproducibility by applying the ideas to replication of a study on organ donation (Kessler and Roth 2012). The target population required to support the statement in the abstract '…significant impact on registration …' would be all people who could potentially register as organ donors, in any country. The original study subjects were students from Boston, USA, who made decisions on whether to donate abstract units named "A" and "B" under various cost structures on computers.

Next, a decision must be made on which effect size (or sizes) to reproduce. The main table of results has over 35 regression coefficients, with robust standard errors, from four ordinary least squares models for binary outcome data (Kessler and Roth 2012 Table 3). Some coefficients are significant. A footnote to Table 3 mentions that multiple other models were fitted. Before Table 3, the authors discuss results based on 'probit tests', giving only p-values in footnotes. Coherent reasons for a choice of study population and particular effects sizes to use in assessing reproducibility

[1]https://www.fda.gov/aboutfda/innovation/ucm512503.htm Accessed 20 July 2018

are left as an exercise for the reader. The challenge appears over-whelming to the current authors.

It is worth reflecting on differences between the statistics and behavioural science communities regarding the treatment of one particular source of variation: people. To statisticians, people are different and represent a source of variation (Altman 1991; Clarke and Kempson 1996). For example, Hand (2004, 152) states that, 'whereas, in physics *all* electrons are *identical*, in psychology *no* two people are identical'. However, to most behavioural scientists, 'there is little variation across human populations' (Henrich, Heine, and Norenzayan 2010b, 61), because 'everyone shares most fundamental cognitive and affective processes, and that findings from one population apply across the board' (Henrich, Heine, and Norenzayan 2010a, 29). It is worth considering the implications of such assumptions on policy decisions and the 'reproducibility crisis'.

2.6 Directions for Future Research

Guidance on reporting standards in psychology has been in place since 1952, and has been continually refined ever since (APA Publications and Communications Board Working Group on Journal Article Reporting Standards 2008). A decade ago, the APA Publications and Communications Board Working Group on Journal Article Reporting Standards (2008, 847) made the following claim:

> *Further, in the spirit of evidence-based decision making that is one impetus for the renewed emphasis on reporting standards, we encourage the empirical examination of the effects that standards have on reporting practices… It can be studied for its effects on the contents of research reports and, most important, its impact on the uses of psychological research by decision makers in various spheres of public and health policy and by scholars seeking to understand the human mind and behaviour.*

Our results suggest that reporting standards have been ignored. We reiterate the call

for work to be done in this nascent but crucial area of research. Some of our future work is already under development to address the effects that standards have on reporting practices. We believe that addressing this question satisfactorily requires close collaboration among various stakeholders, from practitioners, to scholars, funding agencies, learned societies, and journals in the behavioural sciences. It should also closely involve other relevant disciplines, such as medicine and statistics to name but two. However, researchers need to first adhere to recommended practices lest 'the value of the next generation of research may be compromised' (APA Publications and Communications Board Working Group on Journal Article Reporting Standards 2008, 846).

The checklist can also be refined with instructions and shortened to be more practical. In medicine, checklists tend to be shorter and for specific study designs. However, a general statistical practice checklist has value in behavioural science research as there are not many researchers trained as statisticians in the field, especially compared to medicine.

2.7 Conclusion

In the 1990s, the American Psychological Association commissioned a team of psychologists and statisticians to work together, and provide guidance on the use and reporting of statistical methods in psychology journals (Wilkinson and Task Force on Statistical Inference 1999). We adapted their guidelines, and drew on similar guidelines, checklists and recommended practices from other disciplines, to develop a 100-item checklist of recommended statistical practices for economics and psychology research involving human participants.

The results from our evaluation of behavioural science studies provide clear evidence of a failure to report essential information, and adhere to recommended statistical practices. The quality of reporting needs to be improved to facilitate more efficient replication efforts, and ensure that findings are of acceptable statistical standards to influence policy. Behavioural scientists and statisticians should consider work-

ing together again to extend the efforts of Wilkinson and Task Force on Statistical Inference (1999), and possibly refine our checklist. Drawing from the history of medicine, it takes significant time and resources to develop and refine checklists and recommended guidelines (Altman and Simera 2016). An open discussion, based on our detailed findings, involving behavioural scientists and statisticians would be a meaningful step forward.

Is there a reason for the statistics community to get involved, and how might this be achieved? Peng (2015, 32) claims that statisticians 'have an opportunity to attack the crisis of scientific reproducibility at its source' through education and sharing of 'evidence-based data analysis practices'. It is worth considering a quote from 35 years ago by Goldstein (1984, 264):

> *My main thesis is twofold. First, I believe that quantitative social science has everything to gain from an increasing involvement of statisticians. Secondly, I believe that this should involve statisticians coming to regard themselves also as being social scientists.*

Will behavioural scientists embrace close collaboration with statisticians, and vice versa? The well-being of citizens around the world is at stake.

Chapter 3

The Sign Effect in Intertemporal Choice

3.1 Introduction

A major aim of this thesis is to conduct the first systematic review and meta-analysis of the sign effect intertemporal choice anomaly. The previous chapter investigated the standards of statistical practice in behavioural science. The checklist of recommended statistical practice developed in the previous chapter can be used in a systematic review. This chapter introduces the sign effect anomaly. It will attempt to formalise the definition of the sign effect with probabilities and also critically discuss issues concerning, and assumptions of, the sign effect.

In economics, any decision with outcomes that occur at different points in time is known as an 'intertemporal choice'. Read (2004) provides the following examples of intertemporal choices:

- Whether or not to have a flu shot;
- The choice between fruit salad or tiramisu;
- When to get down to work on a promised paper;

- Whether to invest in a pension plan or buy a widescreen TV; and
- (For a pigeon) one food pellet in one second, or two pellets in two seconds.

These examples involve choosing between an earlier and usually smaller penalty or reward (e.g. a flu shot, a TV) and a later and usually larger one (the flu, comfortable retirement). According to Read (2004), the goal of intertemporal choice research is to understand how these choices are made, and how they should be made.

Although there are many examples of intertemporal choices in every day life, intertemporal choice research usually takes place in human experimental laboratories. Participants are recruited and usually presented with a series of binary choice questions involving sooner vs. later amounts of money. An example of one such question might be, 'Would you prefer £100 today or £110 in a year?'

It is common practice to calculate a discount rate for each participant. This discount rate is calculated by inferring an indifference point based on the responses to the binary choice questions. Consider the following example choices:

- Question 1. 'Would you prefer to gain £100 today or gain £110 in a year?'
- Question 2. 'Would you prefer to gain £100 today or gain £120 in a year?'

If a participant chose to receive £100 today in Question 1, and £120 in a year in Question 2, then this implies that the participant is indifferent between receiving £100 today and some amount between £110 and £120 in a year. For the purpose of convenience, researchers may assume the indifference point to be the mid point (Hardisty et al. 2013), which would be £115 in this example.

The indifference point is then used to calculate a discount rate using one or more discounting models, which are mathematical functions. Two of the most popular models used in the literature are the exponential discounting model, and hyperbolic discounting model (Mazur 1987).

The exponential discounting model is given by:

$$V = A\mathrm{e}^{-k_{\mathrm{EXP}}D} \tag{3.1}$$

where V is the present value, A is the future amount, e is the base of the natural logarithm, D is the delay in years, and k_{EXP} is the discount rate.

This can be rearranged to calculate the discount rate, k_{EXP}:

$$V = A\mathrm{e}^{-k_{\mathrm{EXP}}D} \iff \frac{V}{A} = e^{-k_{\mathrm{EXP}}D} \iff ln\left(\frac{V}{A}\right) = -k_{\mathrm{EXP}}D \iff k_{\mathrm{EXP}} = \frac{lnA - lnV}{D}$$

Returning to the previous example, V would be 100, A would be the indifference point of 115, and D would be 1. This would result in k_{EXP} being 0.14. This means that the participant's discount rate is 14%.

The hyperbolic discounting model by Mazur (1987) is given by:

$$V = \frac{A}{1 + k_{\mathrm{HYP}}D} \tag{3.2}$$

where V is the present value, A is the future amount, D is the delay in years, and k is the discount rate.

The hyperbolic discounting model in Equation (3.2) can be rearranged to calculate the discount rate, k_{HYP}:

$$k_{\mathrm{HYP}} = \frac{A}{VD} - \frac{1}{D}$$

Returning to the previous example, V would be 100, A would be the indifference point of 115, and D would be 1. This would result in k_{HYP} being 0.15. This means that the participant's discount rate is 15%.

3.2 The sign effect anomaly

3.2.1 Overview and terminology

The sign effect anomaly is usually described as occurring when 'the discount rate for gains is much higher than for losses' (Loewenstein and Thaler 1989, 187). It is usually tested by showing participants pairs of questions that are similar in all respects except for a switch in the sign. This is an example of a question pair that is similar in all respects except for the sign:

- Question 1a: 'Would you prefer to *gain* £100 today or *gain* £110 in a year?'
- Question 1b: 'Would you prefer to *lose* £100 today or *lose* £110 in a year?'

In the context of binary choice questions, a gain is discounted when the sooner option is chosen while a loss is discounted when the later option is chosen (Hardisty et al. 2013).

There are at least two ways to understand the sign effect. The first is through the concept of opportunity cost. The second is through the concept of discounting.

3.2.1.1 Sign effect from an opportunity cost approach

This approach views all costs (and gains) as opportunity costs (and gains). Consider questions 1a and 1b from the example above. This can be represented in Table 3.2.

Responses to a question pair can then be represented in a 2×2 contingency table, as seen in Table 3.2. There is evidence for the sign effect when $b > c$. Using questions 1a and 1b as an example, cell b represents participants who chose to gain £100 in Question 1a and lose £100 in Question 1b. Cell c represents participants who chose to gain £110 in Question 1a and lose £100 in Question 1b.

Table 3.1: The opportunity cost approach to understand the sign effect using questions 1a and 1b as an example.

Domain	Choice	Now	1 year	Opportunity cost interpretation
Gain (+)				
	'Smaller-sooner'	100		Gain 100 now, lose 110 later
	'Larger-later'		110	Lose 100 now, gain 110 later
Loss (−)				
	'Smaller-sooner'		\$-\$110	Gain 100 now, lose 110 later
	'Larger-later'	\$-\$100		Lose 100 now, gain 110 later

Table 3.2: Illustration of responses to a question pair in a 2 × 2 table using an opportunity cost approach.

	'Smaller-sooner' (-)	'Larger-later' (-)
'Smaller-sooner' (+)	a	b
'Larger-later' (+)	c	d

3.2.1.2 Sign effect from a discounting approach

In the discounting approach, a gain is discounted when the sooner option is chosen while a loss is discounted when the later option is chosen. Consider the previous example of questions 1a and 1b. In Question 1a, the sooner gain is £100 (today) and the later gain is £110 (in 1 year). In Question 1b, the sooner loss is −£100 (today) and the later loss is −£110 (in 1 year).

Let the sooner gain (S^+) and later loss (L^-) be denoted by indicator functions:

$$S^+ = \begin{cases} 1 & \text{if the sooner gain is chosen} \\ 0 & \text{if the later gain is chosen} \end{cases}$$

$$L^- = \begin{cases} 1 & \text{if the later loss is chosen} \\ 0 & \text{if the sooner loss is chosen} \end{cases}$$

Then each participant's response to a given question pair can be represented in Table 3.3. Then, let $X_{i,j}$ be the case when participant i discounts the gain only, i.e. chooses the sooner gain and sooner loss, on question pair j and $Y_{i,j}$ be the case when participant i discounts the loss only, i.e. chooses the later gain and later loss, on question pair j. There is evidence for the sign effect when for the question pair j $\sum_i^N X_{i,j} > \sum_i^N \mid Y_{i,j} \mid$ for $i = 1, 2, ..., N, j = 1, 2, ...J$.

Table 3.3: Individual participant responses to a question pair using a discounting approach.

Discounted?		$S^+ - L^-$	Discounting interpretation
Gain	Loss		
Yes	Yes	1 - 1 = 0	Gain and loss discounted
Yes	No	1 - 0 = 1	Gain discounted only
No	Yes	0 - 1 = -1	Loss discounted only
No	No	0 - 0 = 0	No discounting

3.2.1.3 The link between the opportunity cost and discounting approaches

Both approaches compare one pair of choices with another and produce equivalent outcomes. Consider the same example of questions 1a and 1b. For this question pair, there is evidence for the sign effect when the number of participants choosing to gain £100 *and* lose £100 is greater than the number of participants choosing to gain £110 *and* lose £110.

3.2.2 Illustration

In developed countries, most people use banks to hold and borrow money. Suppose there was a bank offering $p\%$ interest for putting in a deposit, and a building society offering $q\%$ interest on taking a mortgage, with $p \geq 0, q \geq 0$. Assume that consumers are rational and numerate.

If I put in £1,000 as a deposit now and hold it for a year, the value of the deposit will be £1,000 $\times e^{pt}$, where t is the duration of the deposit in years.

If I borrow in £1,000 for a mortgage now and hold the loan for a year, then I can expect to pay back £1,000 $\times e^{pt}$, where t is the duration of the loan in years.

With the deposit interest rate p fixed as $p\% = 2\%$ per year, consider the following three scenarios.

1. If $p > q$, certain gain: at the end of the year, my deposit is worth £1,000 $\times e^{0.02}$, i.e. £1,020.20, and I owe £1,000 $\times e^{0.01}$, i.e. £1,010.10.
2. If $p = q$, neutral.
3. If $p < q$, I have lost money as I pay more to borrow money than I receive by depositing or investing it.

As managers of banks and building societies are numerate and rational, the sign effects hold in commercial transactions.

3.2.3 Issues

There are several issues concerning the sign effect anomaly that need addressing.

3.2.3.1 Imprecise language

The language that has been used to express the concept of discounting for gains and losses in the literature requires greater clarity for two reasons. First, the verbal

descriptions of discounting for losses imply that the absolute value of the monetary loss is being discounted, i.e. $|-£100| < |-£110|$. However, choosing the larger-later loss is an act of discounting because in mathematics: $-100 > -110$. Second, because \$-\$100 is greater than \$-\$110, there is no 'smaller-sooner' or 'larger-later' outcome for losses. It should be a comparison between losing a larger-sooner (e.g. $-£100$ today) or smaller-later (e.g. $-£110$ in one year) outcome.

The imprecise language is also reflected in the lack of a standardised formal definition of the sign effect. The verbal description is broad and leaves room for different interpretations. It may be difficult to compare evidence across studies if they end up addressing different questions or the same questions with different approaches. Greater clarity can be achieved by formalising the definition and terminology in the language of mathematics.

3.2.3.2 One or two sided tests of significance?

> That behaviour is qualitatively different for gains and losses comes as no surprise… Since failure to wait for a reward creates an opportunity cost while postponing a loss incurs an out-of-pocket cost it should be expected that implicit discount rates will be higher for gains as we observe. — Thaler (1981, 206)

Thaler (1981) is credited as the first to demonstrate empirical evidence of the sign effect anomaly in the behavioural science literature. He provides an economic reason for why discount rates for gains should be higher than losses, and not the other way round. This implies that the sign effect anomaly requires a one-sided significance hypothesis test. In a null hypothesis testing framework, one could construct the following hypotheses that attempts to make explicit the verbal definition proposed by Thaler (1981) and colleagues.

Let $k^+(D, A, V)$ be a person's discount rate for gains (e.g. in Q1a), and $k^-(D, A, V)$ be the same person's discount rate for losses (e.g. in Q1b), for particular values of D, A, V. The discount rates have to be estimated from the data. One interpretation

of the sign effect requires: $k^+(D, A, V) > k^-(D, A, V)$. The magnitude of 'much higher than', in the verbal description of the sign effect, is open to interpretation. If the anomaly is interpreted as an average, rather than an individual effect, an alternative interpretation is:

$$\text{sign effect} = \frac{1}{n}\sum_{i=1}^{n} k_i^+(D, A, V) \;>\; \frac{1}{n}\sum_{i=1}^{n} k_i^-(D, A, V) \tag{3.3}$$

where n is the total number of individuals, $k_i^+(D, A, V)$, and $k_i^-(D, A, V)$ represent the respective discount rate for gains, and losses for the i^{th} individual for a given time delay D, present value V, and future amount A.

It is often implicitly assumed that $k_i(D, A, V)$ is constant for an individual, $\forall D, A, V$. We will use as our underlying assumption that a person has their own discount rates, $k_i^+(D, A, V, \underline{X})$ and $k_i^-(D, A, V, \underline{X})$. That is, at the the most general, we allow the discount rates for gain and loss to vary as D, A, V vary, and to depend on a person's characteristics such as age, sex, education, and wealth, denoted $\underline{X}$, a vector of covariates. For example, a meta-analysis of drug treatments for epilepsy show that the effectiveness of CBZ and VPS depend on the patients' age, and type of epilepsy (Cowling et al. 2007).

Then, the null and alternative hypotheses can be expressed as follows:

$$k^+(D, A, V) > k^-(DAV) \;\; \forall \;\; D \geq 0,\; A > 0,\; V \geq 0,\; A \geq V \tag{3.4}$$

The discount rate, k, is continuous and can in theory take on any possible value, i.e. $k \in [-\infty, +\infty]$, but usually $k \in (0, +\infty]$. Sometimes, discount rates are capped due to statistical decisions made in a study, e.g. $k \in [-2.25, +2.25]$ (Hardisty and Weber 2009). It is reasonable to assume that k can be modelled as being generated from a normal distribution, i.e. $k \sim \mathcal{N}(\mu, \sigma^2)$, or a log-normal distribution, i.e. $\ln(k) \sim \mathcal{N}(\mu, \sigma^2)$.

The "expected" one-direction hypothesis test also assumes that people are numerate, and rational in the economic sense of calculating various costs and benefits, e.g. op-

portunity and out-of-pocket costs. A one sided hypothesis test is justified if we are not interested in the possibility that the average discount rate for gains is lower than for losses. Although it may not be "expected", it is still *possible*, particularly if contextual information is not considered.

It is worth considering the following the following quotation on the use of one sided tests in medical research:

> Expectation of a difference in particular direction is not adequate justification (for a one sided test). In medicine, things do not always work out as expected, and researchers may be surprised by their results… If a new treatment kills a lot of patients we should not simply abandon it; we should ask why this happened. —Bland and Altman (1994, 248)

A two-tailed test could be more appropriate to test for the sign effect. If discount rates are estimated because they can be translated for policy purposes, then it should be of interest to understand if and when, on average, people exhibit greater discounting for losses than gains as well. If this possibility were not accounted for in the hypothesis test, and not captured if it did occur, then policy that uses discount rates from academic research might be inaccurate, and thus potentially harmful to citizens.

3.2.3.3 Estimating the discount rate

There are many studies that estimate participants' discount rates to test for the sign effect. The usual approach is to estimate a discount rate for each individual, and then calculate an average of all the estimated discount rates. There are several issues with this approach.

First, the accuracy of the estimated discount rate depends on the accuracy of the estimated indifference point for each individual. If the estimated indifference point is inaccurate, so too will be the estimated discount rate. Good study design is required to ensure the amounts and delay presented in the binary choice questions allow for

accurate interpolation, i.e. with the smallest difference possible between the values at which the choice swaps.

Second, there is an assumption that an indifference point can be estimated for each individual. An indifference point cannot be estimated for individuals who are consistent with choosing either the smaller or larger amount on every question, or if they "switch" their preferences more than once. A discount rate cannot be estimated for individuals who exhibit this behaviour.

There are two common approaches in the literature to work around this issue. The first approach is to exclude such individuals from the analysis, which is considered "standard" practice and 'unavoidable in model-based studies that require successful fits of the (discounting) model' (Lempert, Glimcher, and Phelps 2015, 367). Studies published in prestigious journals have excluded 30% of participants from the analysis (Lempert, Glimcher, and Phelps 2015; Hardisty and Weber 2009). The second approach is to design a study such that 'participants are forced to go back and redo their answers' (Hardisty et al. 2013, 244), thereby ensuring that an indifference point can be estimated.

The third issue with estimating discount rates concerns the analysis. Summarising the estimated discount rates for each participant into a single average discount rate means that variability within and between individuals is not accounted for. The extent of the problem depends on the amount of variation. Next, it is unclear at which unit of analysis should the discount rate be calculated. For example, if the study design had participants engage in several different conditions, then an individual's discount rate can either be estimated for each condition or as one summary measure across all conditions. If the former option was chosen, then the estimated discount rates for each participant are not mutually exclusive and must be accounted for in subsequent analyses. If a single discount rate were calculated, then information and variability are lost.

3.3 Formalising the sign effect with probabilities

3.3.1 Notation

The conventional approach to studying the sign effect has been to use an economic discounting framework, where discount rates are estimated for participants in a study. An alternative approach is to formalise the sign effect in probabilities. We attempt to do so in this section by first setting out the required notation. For simplicity, dependence on $A,\ V,\ \underline{X}$ is suppressed.

Notation for gains. Let S^+ be a random variable, such that:

$$S^+ = \begin{cases} 1 & \text{if the sooner gain is chosen} \\ 0 & \text{if the later gain is chosen} \end{cases} \tag{3.5}$$

We have random variables $S^+(k^+, D, A, V, \underline{X})$, which under our assumptions for gains:

$$P(S^+ = 1 \mid k^+, D, A, V, \underline{X}) = \begin{cases} 1 & V > Ae^{-k^+D} \\ 0 & V < Ae^{-k^+D} \end{cases} \tag{3.6}$$

$$\Rightarrow P(S^+ = 1 \mid k^+, D, A, V, \underline{X}) = I_{V > Ae^{-k^+D}} \tag{3.7}$$

Example. An individual is asked to choose between receiving £100 today or £110 in a year (e.g. through a fixed deposit in a bank). If the individual has a discount rate of 2%, i.e. $k^+_{i=1} = 0.02$, then using the exponential discounting model in Equation (3.1), the future amount (of £110 in a year) is discounted and valued at £104.60 today. The £100 offered today is worth less than £104.60, i.e. $V < Ae^{-k^+D}$, and so the individual chooses to wait and take the later gain, i.e. $S^+ = 0$. If a second individual has higher a discount rate of 20%, i.e. $k^+_{i=2} = 0.2$, then using the same model, the future amount is worth £90.10 today. The £100 offered today is worth

more than £90.10, i.e. $V > Ae^{-k^+D}$, and so the second individual chooses to take the sooner gain, i.e. $S^+ = 1$.

The notation for losses is as follows. Let S^- be a random variable, such that:

$$S^- = \begin{cases} 1 & \text{if the sooner loss is chosen} \\ 0 & \text{if the later loss is chosen} \end{cases} \tag{3.8}$$

We have random variables $S^-(k^+, D, A, V, \underline{X})$, which under our assumptions for losses:

$$P(S^- = 1 \mid k^-, D, A, V, \underline{X}) = \begin{cases} 1 & -V > -Ae^{-k^-D} \\ 0 & -V < -Ae^{-k^-D} \end{cases} \tag{3.9}$$

$$\Rightarrow P(S^- = 1 \mid k^-, D, A, V, \underline{X}) = I_{-V > -Ae^{-k^-D}} \tag{3.10}$$

Example. An individual is asked to choose between paying £100 today or £110 in a year (e.g. through a loan from a bank). If the individual has a discount rate of 2%, i.e. $k^-_{i=1} = 0.02$, then using the exponential discounting model in Equation (3.1), the future amount (of −£110 in a year) is discounted and valued at -£107.80 today. Since −£100 > −£107.8, i.e. $-V > -Ae^{-k^-D}$, the individual chooses the sooner loss, i.e. $S^- = 1$. Similarly, a second individual with a higher discount rate of 20%, i.e. $k^-_{i=2} = 0.2$, would value the future loss at −£90. today. Since −£100 < −£90, i.e. $-V < -Ae^{-k^-D}$, the individual chooses the later loss, i.e. $S^- = 0$.

It may also be worth considering the situation when $V = Ae^{-kD}$, where people are completely indifferent and the probability of choosing either option is 1/2.

3.3.1.1 Hypothesis testing framework

$$\begin{aligned} \mathrm{H}_0 &: k^+ = k^- \quad \forall D, A, V \\ \mathrm{H}_1 &: k^+ > k^- \quad \forall D, A, V \end{aligned} \tag{3.11}$$

This is equivalent to the mathematical definition of the sign effect proposed by Han and Takahashi (2012), who used the 'q-exponential model' to define the 'degree of the sign effect' as:

$$\text{Sign Effect} := \frac{k_q^+ - k_q^-}{k_q^+} \tag{3.12}$$

where k_q^+ and k_q^- are each a 'free parameter' indicating 'the degree to which a subject discounts the delayed' gain and loss respectively based on the 'q-exponential model'. If $\frac{k_q^+ - k_q^-}{k_q^+} = 0$, then there is no sign effect, which is equivalent to H_0 in equation (3.11) as $k_q^+ = k_q^-$. If $\frac{k_q^+ - k_q^-}{k_q^+} > 0$, then there is a sign effect, which is equivalent to H_1 in equation (3.11) as $k_q^+ > k_q^-$.

One possible estimator for an individual i, is an estimate of the probability of choosing gains and losses respectively:

$$\begin{aligned} \overline{S}^+ &= \frac{1}{v}\sum_{j=1}^{v} S_{i,j}^+ \\ \overline{S}^- &= \frac{1}{v}\sum_{j=1}^{v} S_{i,j}^- \end{aligned} \tag{3.13}$$

Where $S^+ \sim \text{Binomial}(n, \pi^+)$, and $S^- \sim \text{Binomial}(n, \pi^-)$.

The definition for the sign effect based on the discount rate, i.e. $k^+ > k^- \forall D,\ A,\ V$, can be rewritten as:

$$\text{sign effect} = \pi^+ - \pi^- > 0 \tag{3.14}$$

Chapter 4

A Systematic Review of the Sign Effect Anomaly

4.1 Study aim

The aim of this study is to conduct the first systematic review and meta-regression analysis of the 'sign effect anomaly'. A systematic review may be defined as 'a review of a clearly formulated question that attempts to minimise bias using systematic and explicit methods to identify, select, critically appraise and summarise relevant research' (Siddaway, Wood, and Hedges 2018; Needleman 2002, 6). A meta-regression analysis is a regression analysis that combines results from various studies to 'relate the size of the effect to one or more characteristics of the studies involved' (Thompson and Higgins 2002, 1559).

This quantitative review will focus on studies using binary choice question pairs involving monetary gains and losses. Binary choice questions are very commonly used in the intertemporal choice literature. A question pair is two questions that are similar in all respects except with a switch in their signs An example of a question pair is:

Question 1a. 'Would you prefer to gain £100 today or gain £110 in a year?'

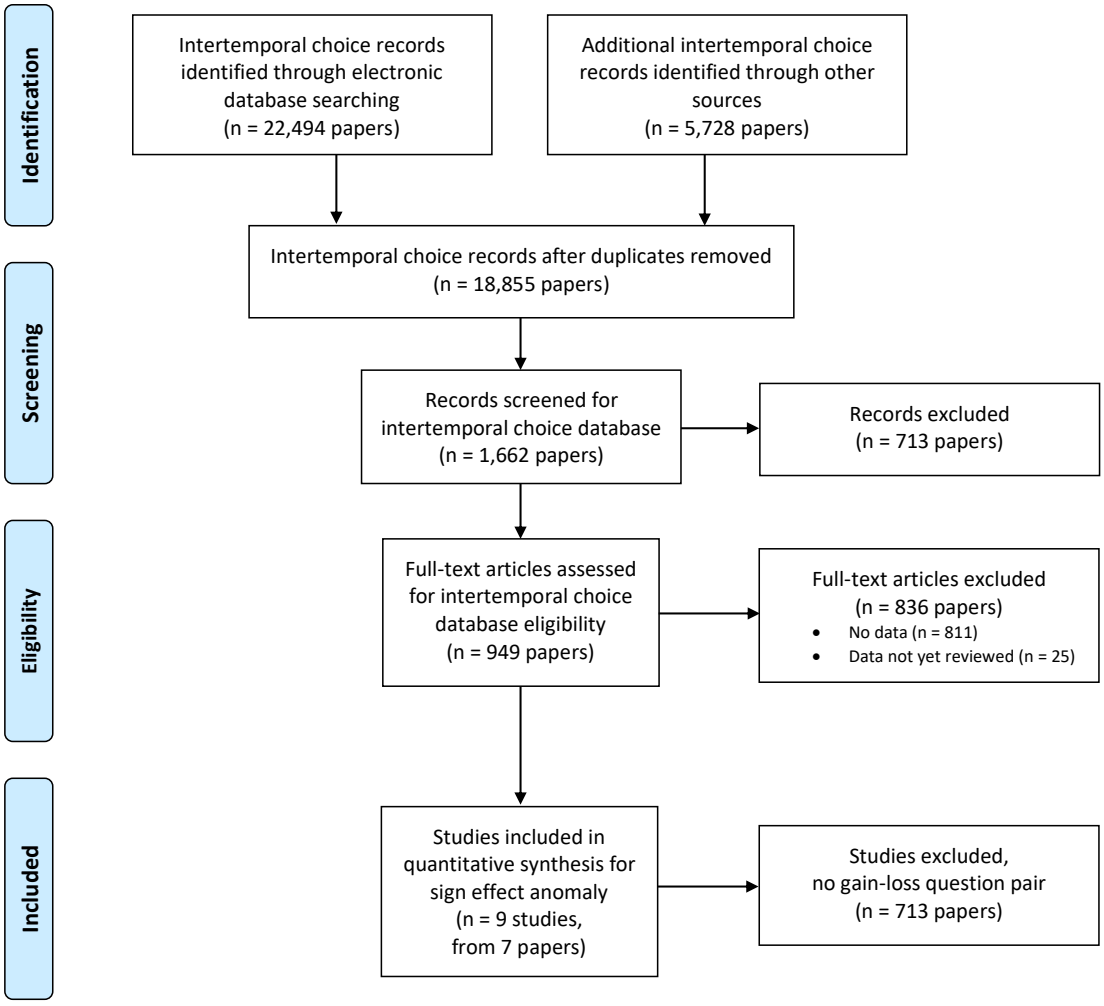
Question 1b. 'Would you prefer to lose £100 today or lose £110 in a year?'

This is the outline of the chapter. Section 4.2 describes the data source and major assumptions made when transforming data received from study authors. Section 4.3 provides descriptive statistics based on data aggregated at the question-level. It looks at the relationship between choosing the later option with sign, amount ratio and time delay differences, which are common factors across all studies. Section 4.4 focusses specifically on the sign effect. It provides a definition of the sign effect, explores the anomaly at the question- and individual-level, challenges assumptions based on the data, and highlights the difficulties in analysing the sign effect in its current definition.

4.2 Data source

The data were from a database of intertemporal choices involving sooner vs. later amounts of money. KTKL, LH, and DR collected the data for a separate project and identified 949 suitable studies for the database. They requested data from all authors via email for the database and were able to include 113 studies. Each observation in the database represents the aggregated response, i.e. the proportion of participants choosing the later option for one question within a study. All studies that presented gain and loss question pairs, regardless of study design, were eligible for inclusion in this chapter. This may include studies that do not explicitly report testing for the sign effect. The entire process is shown in the PRISMA flowchart in Figure 4.1.

Some authors provided the raw data of each choice each participant made to each question. Other authors provided each individual participant's indifference points for a given set of covariates based on the study design, e.g. for each unique combination of delay and money amounts. The indifference points for each individual was then used as an indicator for when the participant switched choices in order to calculate each choice made. This calculation method assumes that each participant has only one switching point for a given set of unique combination of covariates.



Adapted from: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). *P*referred *R*eporting *I*tems for *S*ystematic Reviews and *M*eta-*A*nalyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097

Figure 4.1: PRISMA flowchart of studies included in this chapter.

For example, only indifference points were available for Hardisty et al. (2013) and Han and Takahashi (2012). Hardisty et al. (2013) varied the later amounts and later delay while keeping the sooner amount constant at USD300. A script was written to calculate individual choices, by comparing a participant's indifference point at each delay with each value of the later amount. If the later amount was greater than the indifference point, then the participant chose the later amount for gains. For losses, if the indifference point of a participant was greater than the later amount, then the participant chose the later amount.

Han and Takahashi (2012) varied the sooner amounts and later delay while keeping the later amount constant at JPY100,000. Each participant's indifference point for each later delay was compared with each value of the sooner amount. If the indifference point was greater than the sooner amount, then the participant chose the later amount for gains. For losses, if the indifference point was smaller than the sooner amount, then the participant chose the later amount.

For example, in Han and Takahashi (2012), the sooner amount ranged from JPY0 to JPY100,000, with an increase of JPY2,500 each time. Suppose a participant had an indifference point of 98,750 for gains and for losses. The script would begin with checking if 98,750 was larger than the smallest sooner amount, i.e. 0, and determine that the participant chose the later amount of JPY100,000 for gains. In this case, the participant would continue to choose the later gain until the largest sooner amount, where the participant switches preferences and chooses the sooner amount of JPY100,000 now.

For losses, the participant would choose the sooner amount on each question and switch to choosing the later loss for the maximum value of the sooner amount, i.e. when the sooner and later amounts are equal. If the same rule for gains were applied for losses, then the participant would always choose the later loss until the last option, which would not make sense. For example, no rational person would choose to lose JPY100,000 in 7 days when they could pay/lose nothing now, i.e. the smallest sooner amount.

Indifference points were available, but individual choices could not be determined for

McKerchar, Pickford, and Robertson (2013) who reported calculating the indifference points at each delay based on the outcome sign and order in which the sooner amount was presented. However, the order was not provided in the data with the indifference points. As such, data from McKerchar, Pickford, and Robertson (2013) could not be used in this chapter.

Finally, the study from Chen10 was reported in Chinese and requires further verification. The data were explored but not used in any formal statistical models.

4.3 Descriptive statistics at the question level

This section provides descriptive statistics of the eligible studies which we have data for. Data at the question-level, i.e. where each observation is the aggregated number of responses to a question asked, are explored. All analysis was done with the R language (version 3.4.3).

4.3.1 Study characteristics

Nine studies from 7 papers met the inclusion criteria, and were extracted from the database. Table 4.1 summarises the important characteristics of the 9 included studies.

There is substantial heterogeneity in the characteristics across studies. There are 1,664 questions/observations across all studies, i.e. 832 question pairs. The number of questions presented to participants range from 3 to 574. The number of participants range from 20 to 118. There are big differences in the amounts of money presented. Most studies present the sooner amount of money today, although the maximum is 184 days. The delay to the larger amount of money range from 7 days to 25 years, i.e. 9,125 days. The implied annual interest rates, which were inferred from the amounts of money available and the delays, range from -30.6% to infinity.

Figure 4.2 shows a strip chart with the amount ratio on the y-axis, and the Study ID on the x-axis. The ratios are used as the amounts can vary greatly within

and between studies, which makes it difficult to display meaningfully on one graph. Each point is jittered horizontally, and represents one observation. There are 832 observations (question pairs). Each solid red line represents the median value for the study. Studies are ordered by increasing median value.

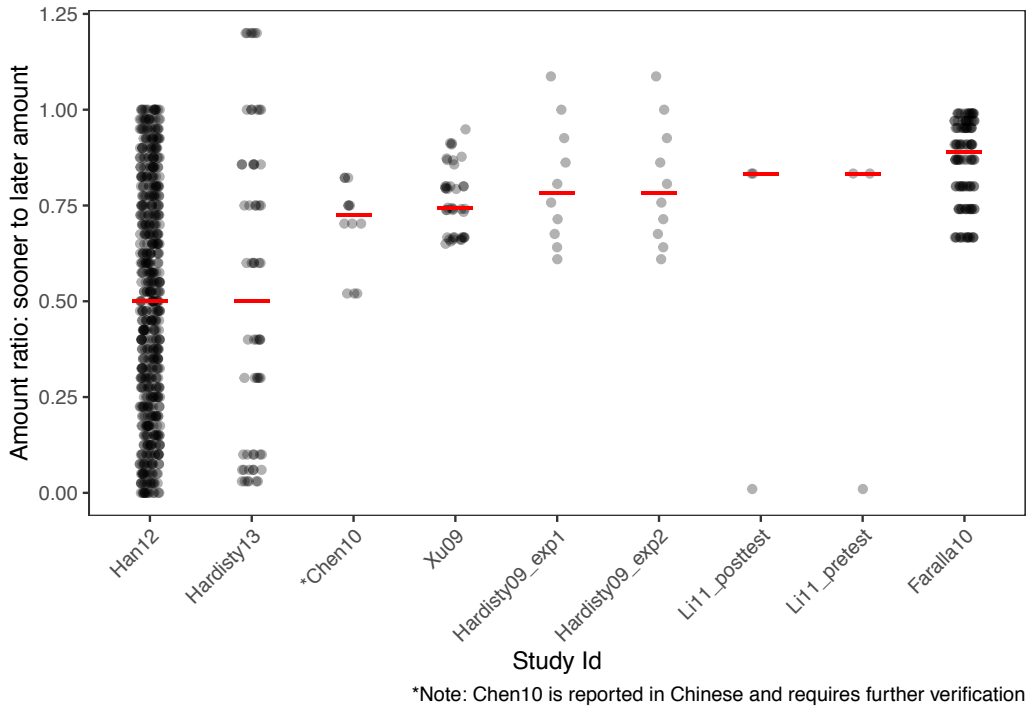


Figure 4.2: Jittered strip chart: Amount ratio by Study Id for 832 question pairs. Amount ratio is defined as the sooner amount divided by the later amount. Studies are ordered by increasing median amount ratio. A solid red line represents the median value for the study.

From Figure 4.2, the first study from the left have observations that are evenly spread from around 0 to 1.0 on the y-axis, with a median value of 0.5. An example of an amount ratio of 0.5 would be a sooner amount of £100 and a later amount of £200. The second study also has a median value of 0.5 but values range from 0 to 1.2. The other 7 studies have median values of 0.7 and higher. Also, observations from the other 7 studies tend to be between 0.6 and 1.0 on the y-axis, with almost no amount ratios below 0.6 in value.

Table 4.1: Summary of characteristics for the 9 included studies from 7 papers. Obs: number of questions each participant responded to. N: number of participants. Sooner and later delays presented as days. Parenthesis indicates the corresponding number of unique values when it is more than 1.

ID	Obs	N	Currency	Sooner amount	Later amount	Sooner delay	Later delay	Annual interest rate
Han12	1,148	49-50 (2)	JPY	0-1e+05 (41)	1e+05	0	7-9,125 (7)	0-Inf (274)
Faralla10	240	25	EUR	5-30 (3)	5.05-45 (24)	0-30 (3)	14-46 (3)	12-3.9e+06 (32)
Hardisty13	120	53-54 (2)	USD	300	250-1e+04 (10)	0	182-3,650 (3)	-30.6-1.1e+05 (28)
Xu09	80	20	CNY	13-110 (31)	20-149 (32)	0-30 (3)	14-61 (4)	185-4.9e+06 (37)
Chen10	24	50	CNY	370-900 (4)	450-1,480 (4)	61-184 (8)	392-1,096 (8)	29.8-3.3e+01 (8)
Hardisty09_exp1	20	65	USD	250	230-410 (10)	0	365	-8-6.4e+01 (10)
Hardisty09_exp2	20	118	USD	250	230-410 (10)	0	365	-8-6.4e+01 (10)
Li11_posttest	6	101-104 (2)	CNY	10-3,000 (3)	120-3,600 (3)	0	365	20-9.9e+03 (2)
Li11_pretest	6	98-104 (3)	CNY	10-3,000 (3)	120-3,600 (3)	0	365	20-9.9e+03 (2)

Figure 4.3 displays the time delay ratio by Study Id for 172 question pairs. The excluded studies always had the sooner amount available today. Each point is jittered horizontally, and represents one observation. Studies are ordered by increasing median value. There are not many different time delay ratios.

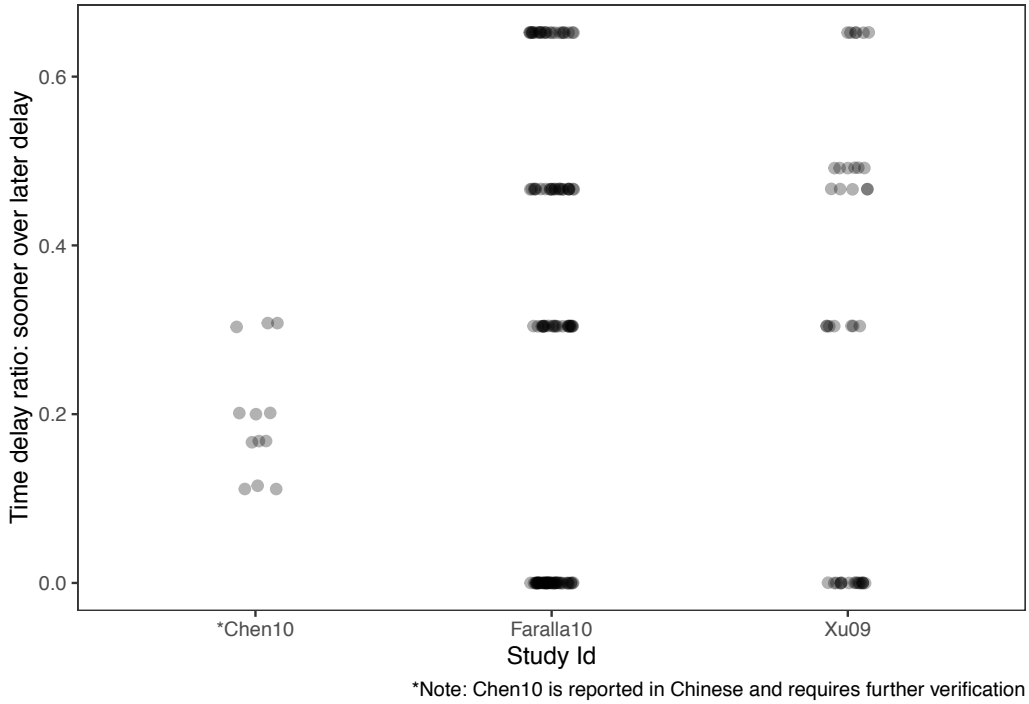


Figure 4.3: Jittered strip chart: Time delay ratio by Study Id for 3 studies. The excluded studies always had the sooner amount available today. The time delay ratio is defined as the sooner delay divided by the later delay. Studies are ordered by increasing median delay ratio. Points are jittered horizontally.

4.3.2 Choosing the later option by sign

Figure 4.4 shows a density plot of the proportion choosing the later option coloured by sign for all 1,664 questions from the 9 studies. Figure 4.5 displays the proportion choosing the later option coloured by sign for each study except Li, which only had 6 questions in each study.

From Figure 4.4, there is a distinct pattern of the proportion of later choices by sign. The density for losses falls mostly below 0.5, while the density for gains is greater

above 0.5. This suggests that there is a greater proportion of later gains and sooner losses. However, there is heterogeneity across studies (Figure 4.5).

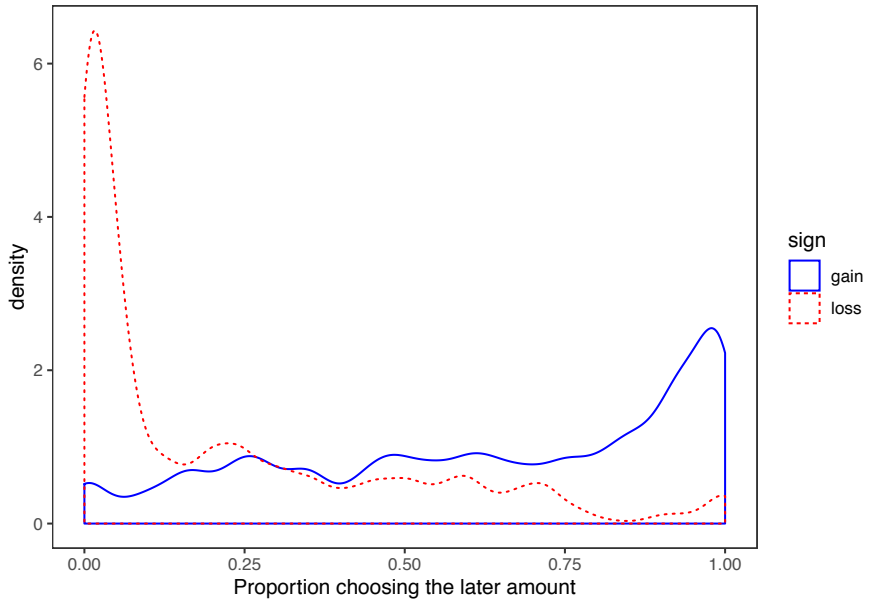


Figure 4.4: Porportion of participants choosing the later amount across all studies.

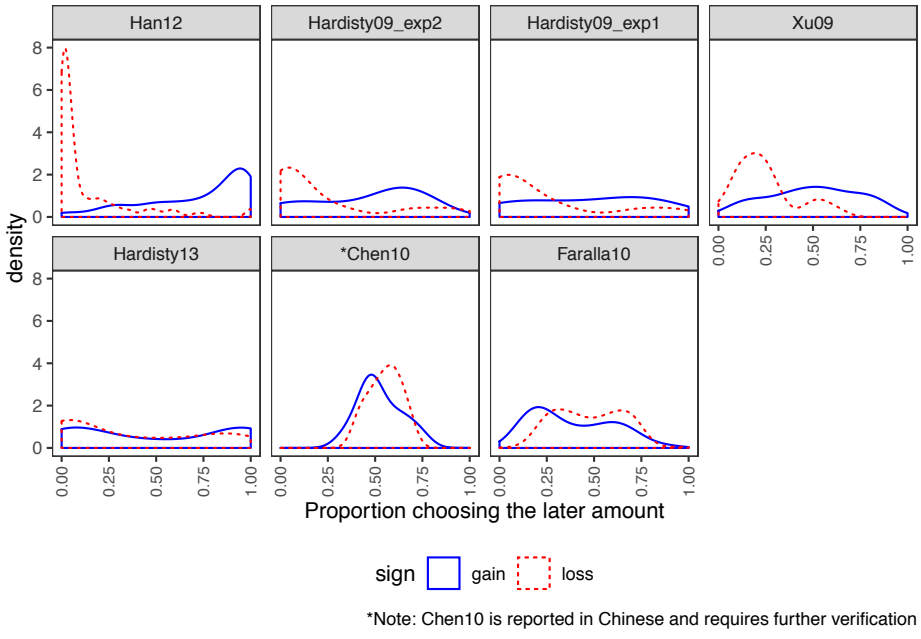


Figure 4.5: Density plots of the proportion of participants choosing the later amount for each study. Two studies, from Li11, are dropped as they each contain 6 observations (3 for gains, 3 for losses).

4.3.3 Choosing the later option and amount ratio

Figure 4.6 displays a scatter plot of the relationship between the proportion choosing the later option and the amount ratio, which is defined as the sooner amount divided by the later amount, for each study. A ratio of zero means the sooner amount was 0 and a ratio of 1.00 means the sooner amount was equivalent to the later amount. The points are coloured by sign.

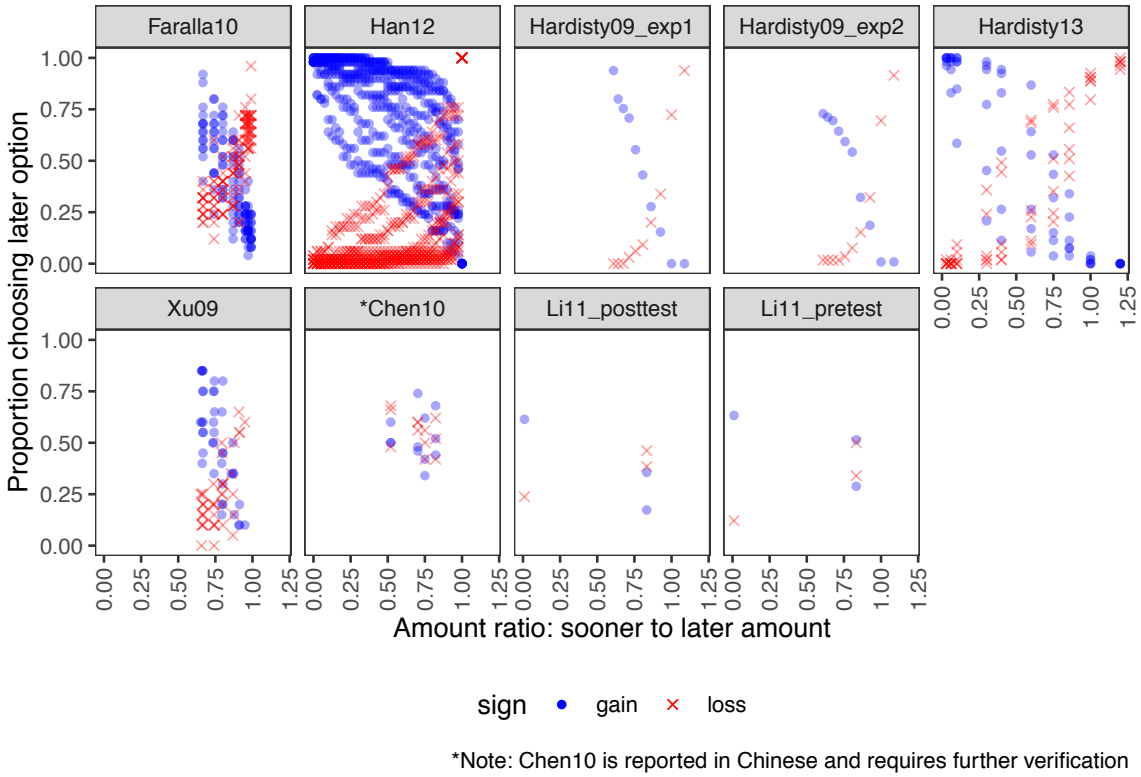


Figure 4.6: Proportion of participants choosing the later amount by amount ratio, coloured by sign. Each point represents the proportion of participants choosing the later amount on a question.

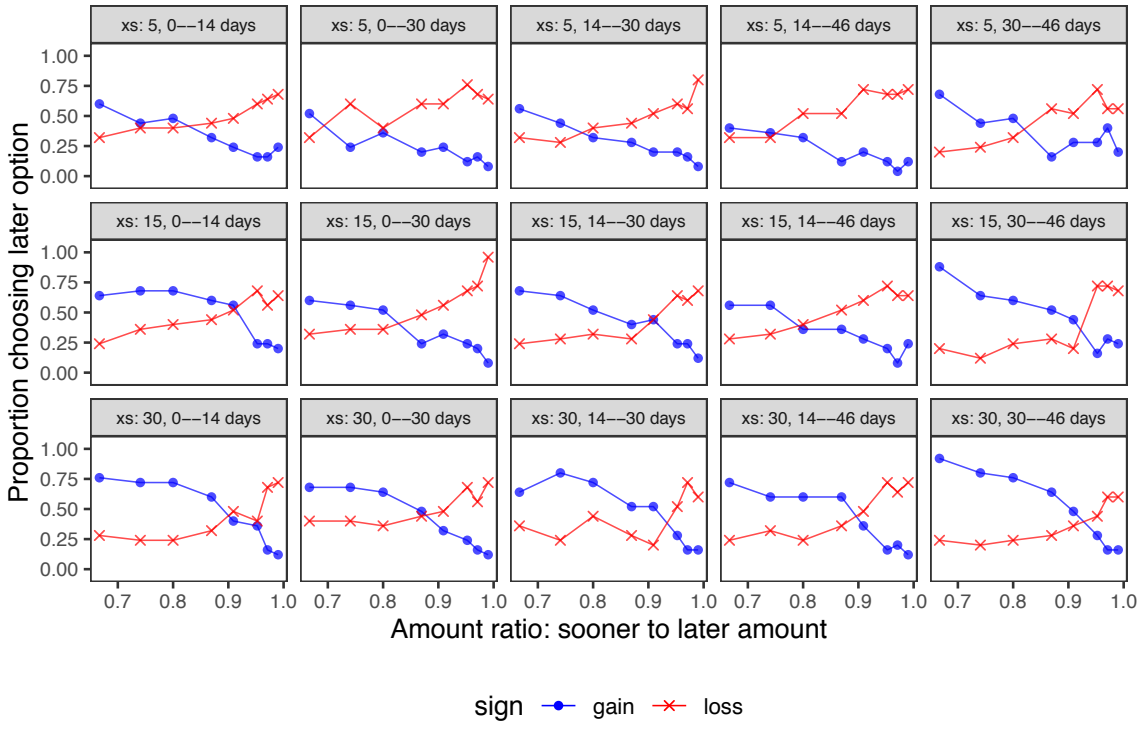
In several studies, the proportion choosing the later gain tends to decrease as the amount ratio increases while the proportion choosing the later loss tends to increase as the amount ratio increase. The rate of change in these relationships varies across studies.

Some studies have within-subject factorial designs such that a relationship can be

displayed for each unique combination of factors. However, the study designs are not always explicitly reported. These are displayed in Figures 4.7, 4.8, 4.9 and 4.10. Each point represents the proportion choosing the later option on one question.

4.3.3.1 Faralla et al. (2012)

This study appears to have a: 3 (sooner amount) × 5 (combination of sooner and later delay) × 2 (sign) factorial design. Each participant was asked 240 questions in total, of which half were gains and half were losses. Figure 4.7 shows the relationship between the proportion choosing the later option by the amount ratio coloured by sign. Generally, as the amount ratio increases, the proportion choosing the later option increases for losses but decreases for gains. However, the rate of change varies across the panels.



Panel headings correspond to the sooner amount, and sooner and later delays, reflecting the study design of Faralla10

Figure 4.7: Proportion choosing later option by amount ratio coloured by sign for Faralla10.

4.3.3.2 Hardisty et al. (2013)

This study appears to have a: 3 (later delay) × 2 (sign) factorial design. Each participant was asked 120 questions in total, of which half were gains and half were losses. Figure 4.8 shows the relationship between the proportion choosing the later option by the amount ratio coloured by sign.

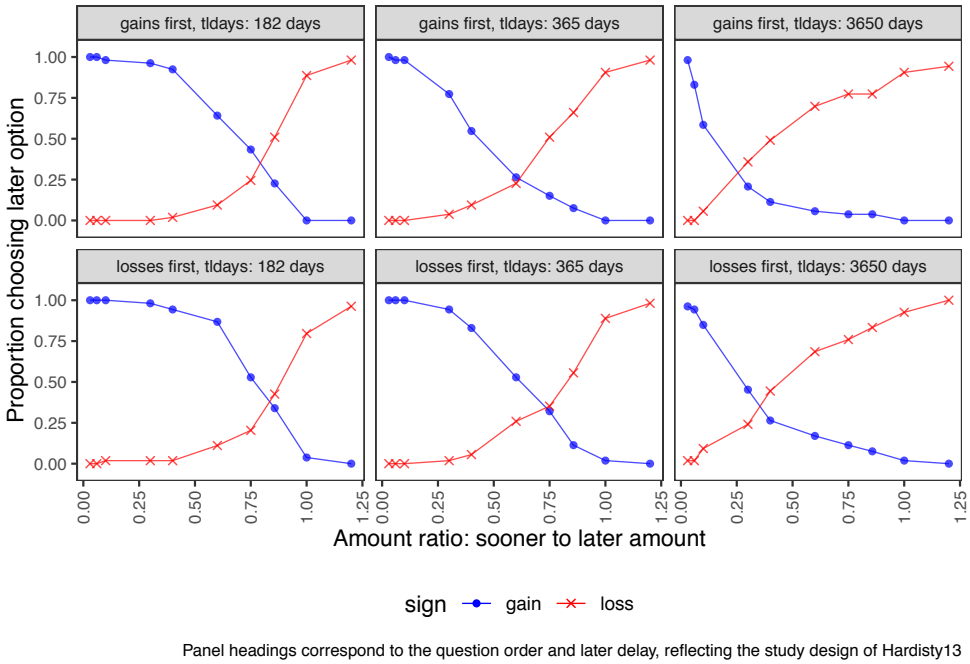


Figure 4.8: Proportion choosing later option by amount ratio coloured by sign for Hardisty13. The panels refer to the different time delays for the later amount.

4.3.3.3 Han and Takahashi (2012)

This study appears to have a: 2 (presentation order) × 7 (later delay) × 2 (sign) factorial design. The presentation order refers to whether participants saw the sooner amounts in an ascending or descending order, where the later amount was always JPY100,000. Each participant was asked 1,148 questions in total, of which half were gains and half were losses. Figure 4.9 shows the relationship between the proportion choosing the later option by the amount ratio coloured by sign.

It is interesting to note the spike in the proportion choosing the later option when

the amount ratio is 1.00. The proportion choosing the later gain tends to 0 while the proportion choosing the later loss tends to 1 when the sooner and later amounts are equal. For example, if participants were asked to choose between £100 today and £100 in 7 days, then almost every one would choose to gain £100 today. Using the same example for losses, almost every participant would choose to lose £100 in 7 days rather than lose £100 today.

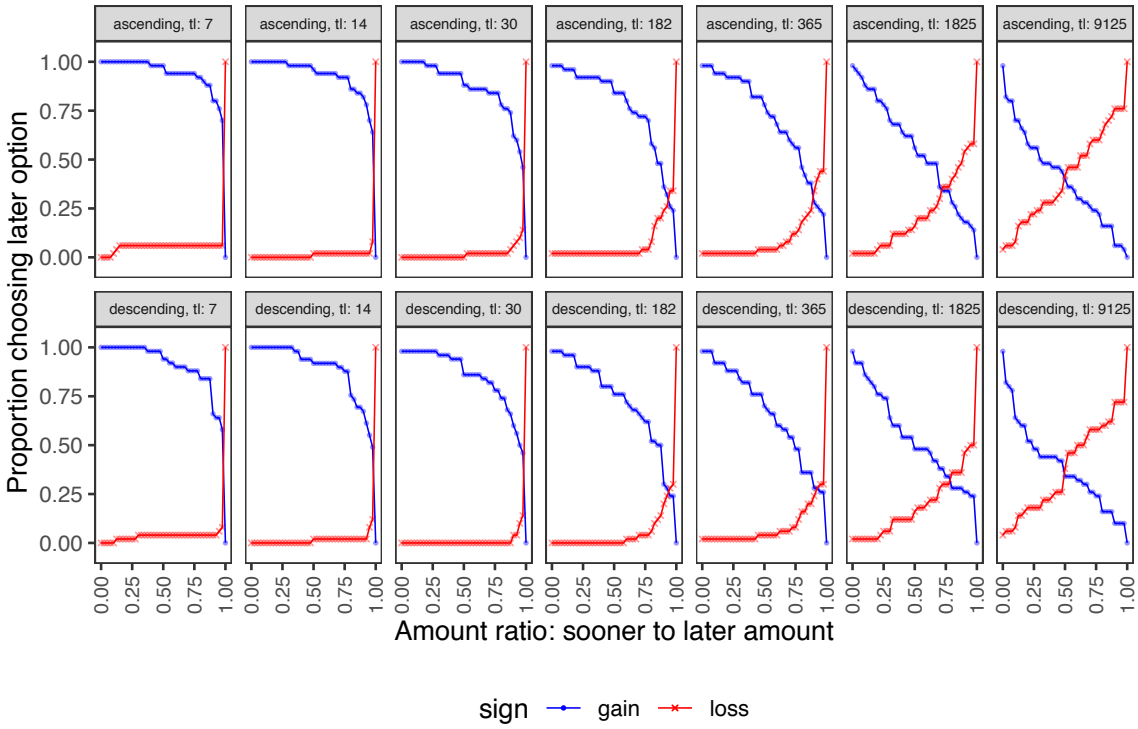


Figure 4.9: Proportion choosing later option by amount ratio coloured by sign for Han12.

4.3.3.4 Xu et al. (2009)

This study appears to have a: 6 (combination of sooner and later delays) × 2 (sign) factorial design. Each participant was asked 80 questions in total, of which half were gains and half were losses. Figure 4.10 shows the relationship between the proportion choosing the later option by the amount ratio coloured by sign.

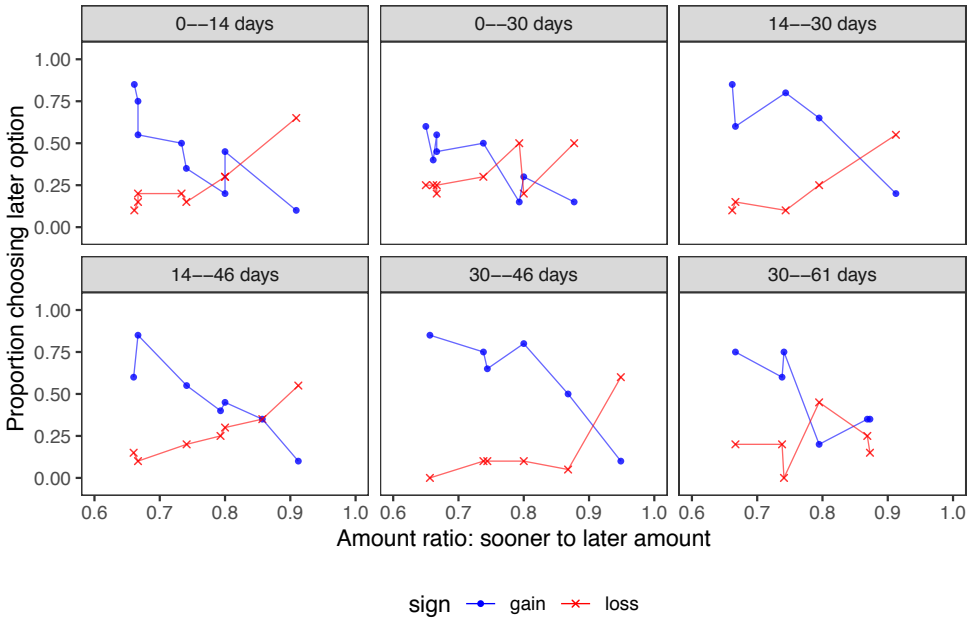


Figure 4.10: Proportion choosing later option by amount ratio coloured by sign for Xu09.

4.3.4 Choosing the later option and time delay difference

Figure 4.11 displays the relationship between the proportion choosing the later option and difference in when the sooner and later options are available. The proportion choosing the later gain tends to decrease, while the proportion choosing the later loss tends to increase, as the time delay difference increases.

4.4 Sign effect analysis

This section begins by defining the sign effect formally. Then, it explores the sign at the question-level first and then the individual-level for each separate study. It disentangles the concept of discounting into mutually exclusive behaviours and examines the frequency of each behaviour. Table 4.2 summarises key characteristics that are related to the sign effect for the 9 included studies.

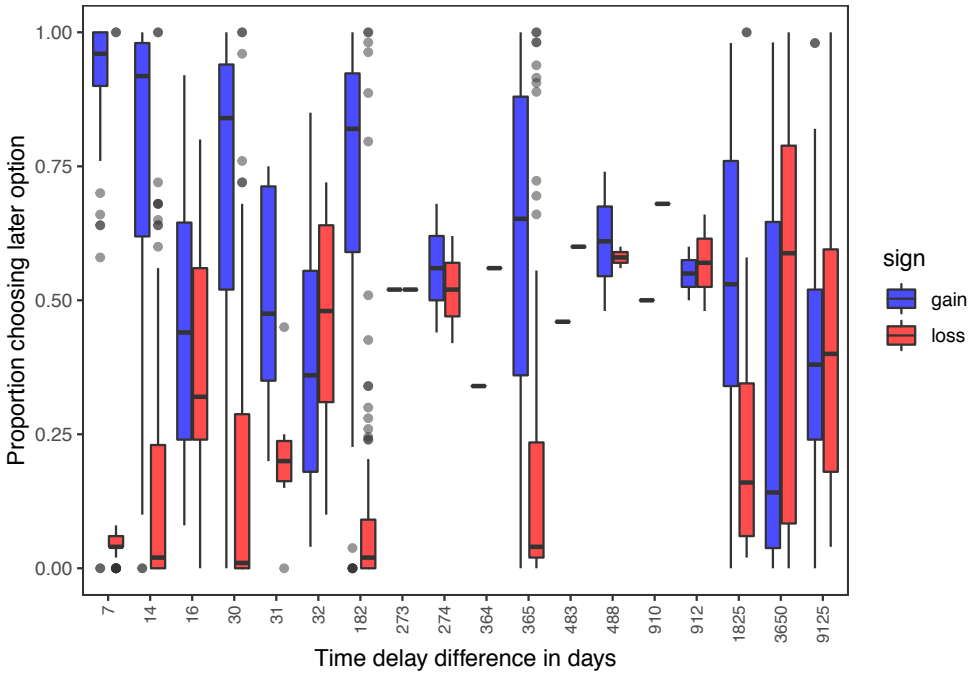


Figure 4.11: Proportion choosing the later option by time delay difference. The delay difference is calculated by subracting the sooner delay from the later delay.

There appears to be no consistent method for analysing the sign effect. Only one study provided a mathematical definition of the sign effect. Most studies provided generic descriptions of their design. It was common for studies to report dropping participants, e.g. Hardisty09 report dropping 16 participants (from a sample of 90) whose responses 'switched back and forth more than once or switched in a manner that would make sense only if they preferred more losses or fewer gains' (Hardisty and Weber 2009, 331). There is also the case of Hardisty13 where 'participants were forced to go back and redo their answers' if their responses indicated 'nonmonotonicity or perverse switching' (Hardisty et al. 2013, 10).

There is also a lack of consistent definition or method for analysing the sign effect. We provide proposals for analysing the sign effect at the question and individual level. To analyse the sign effect at the individual level, we use data from the 6 studies that provided individual participant data (IPD).

We first begin with exploratory data analysis before formal statistical modelling.

Table 4.2: Summary of key characteristics related to the sign effect for the 9 included studies from 7 papers. Study designs are displayed as how they were reported. Note: Chen10 is reported in Chinese and requires further verification.

Study	IPD?	Data available	Reported sign effect?	Outcome analysed	Statistical test	Study design reported
Faralla10	Yes	Raw choices	Yes (medium / large amounts)	Sooner/later choices (%) by sign and money amounts	None	Experiment
Xu09	Yes	Raw choices	Yes	Sooner choices (%)	t-test	Experiment
Hardisty09_Exp1	Yes	Raw choices	Yes	Average discount rate	t-test	Factorial
Hardisty09_Exp2	Yes	Raw choices	Yes	Average discount rate	t-test	Experiment
Han12	Yes	Indifference points	Yes	Average discount rate	None	Experiment
Hardisty13	Yes	Indifference points	Yes	Average discount rate	R^2	Experiment
Chen10	No	Question-level	?	?	?	?
Li11_Exp1	No	Question-level	No	Unclear	Logistic regression	Test-retest
Li11_Exp2	No	Question-level	No	Unclear	Logistic regression	Test-retest

Question-level: Proportion of participants choosing later option for each question

4.4.1 Question-level: Exploratory data analysis

The following formula is one way of calculating an effect size for the sign effect anomaly at the question level:

$$\text{es}_{q,j} = \frac{S^{+}_{q,j} - L^{-}_{q,j}}{N_{q,j}} \tag{4.1}$$

where $\text{es}_{q,j}$ is the effect size on the q^{th} question pair in the j^{th} study, $S^{+}_{q,j}$ and $L^{-}_{q,j}$ refer to the number of participants choosing the sooner gain, and later loss respectively for the q^{th} question pair in the j^{th} study, and $N_{q,j}$ is the total number of participants responding to the q^{th} question pair in the j^{th} study.

This effect size is a difference in proportion, which can range from -1 and $+1$ inclusive. A positive effect size would provide evidence for the sign effect. A negative effect size would provide evidence for a reverse sign effect. An effect size of zero would represent no or zero effect.

Figure 4.12 shows the distribution of the effect sizes for the 832 question pairs across 9 studies. A density curve is plotted over a histogram, with a vertical red median line. This does not show between and within study variation.

The effect size ranges from -0.3 to 0.53, with a median value of 0.14. Of the 832 question pairs, 12% have a zero effect size, 7% have a negative effect size, and 82% have a positive effect size.

Figure 4.13 shows a strip chart with the effect size of the sign effect on the y-axis, and the study ID on the x-axis. Each point is jittered horizontally, and represents one observation. There are 832 observations. The red, horizontal dashed line represents the overall median value. The blue, horizontal dotted line represents the zero line of no effect. The solid red lines represent the median effect size for each study. Studies are ordered by increasing median value.

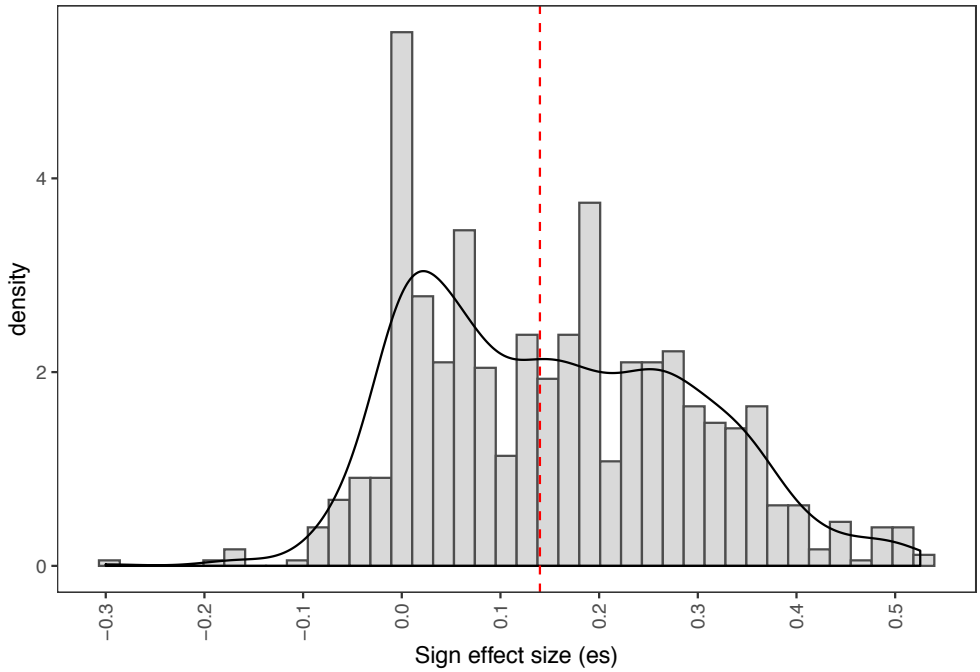


Figure 4.12: Sign effect size for 832 question pairs across 9 studies, with vertical red median line.

From Figure 4.13, there is heterogeneity in the effect sizes across the 9 studies. The median effect size within a study ranged from -0.07 to 0.32 across the 9 studies. There is also heterogeneity in effect sizes within a study as the effect sizes tend to be spread out.

Figure 4.14 shows density plots of the amount ratio on the x-axis, which are coloured by the direction of the sign effect effect size. The red, green, and blue curves represent the amount ratio when the respective effect size is negative, positive, and zero.

There are three distinct patterns in Figure 4.14. The negative red curve has a hump from 0 to 0.5 and a peak at around 0.75. The positive green curve is negatively skewed. The zero effect green curve has a positive skew with two prominent peaks.

Figure 4.15 shows the relationship between the sign effect size on the y-axis, and the amount ratio on the x-axis, for each study. The amount ratio is defined as $\frac{\text{xs}}{\text{xl}}$. Each point is one observation. There are 832 observations in total.

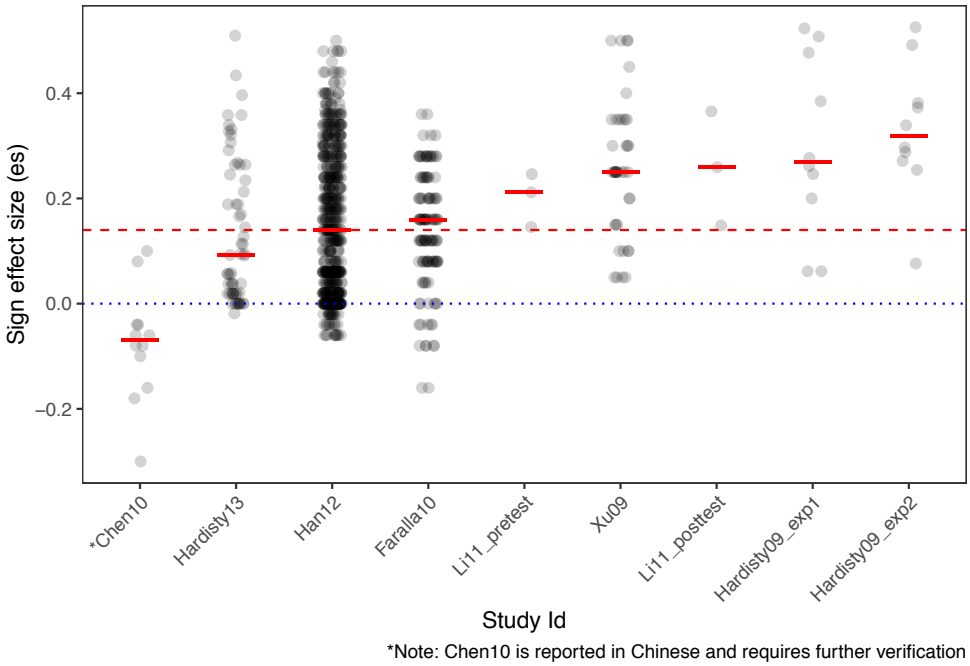


Figure 4.13: Sign effect size for 832 question pairs by study Id. Points are jittered hozitontally. Horizontal dotted blue line represents zero line of no effect. Horizontal dashed red line represents overall median value. Solid red lines represent median value for each study Id.

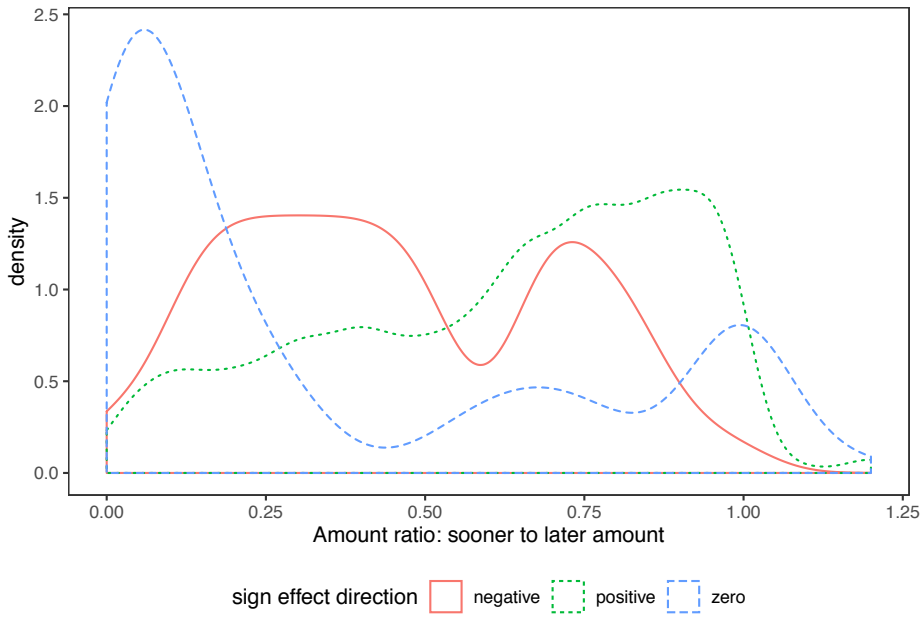


Figure 4.14: Density plot of amount ratio coloured by direction of sign effect size across all 9 studies.

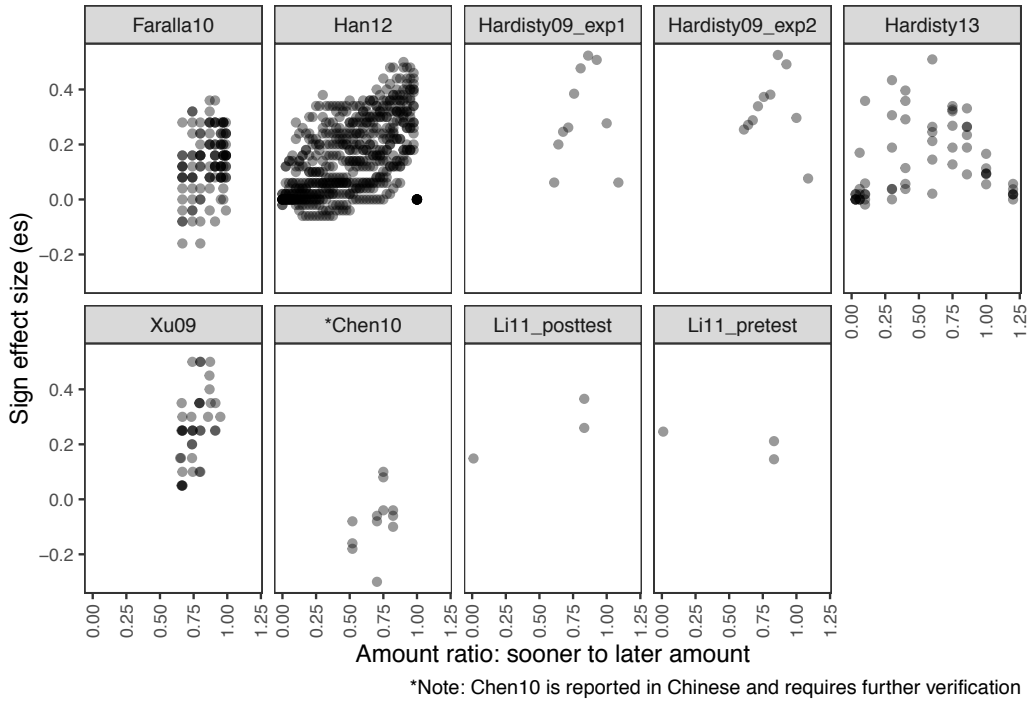


Figure 4.15: Scatter plots of sign effect size by amount ratio for each study.

There is no clear relationship between the sign effect size (es) and the amount ratio in Figure 4.15. Some studies appear to have a positive relationship, e.g. Han12, while others do not appear to show any discernable relationship, e.g. Hardisty09.

Figures 4.16 to 4.19 show this relationship in greater detail for individual studies. The relationship for each study is shown for the factorial combinations in the study design. The approach and rationale are outlined in Section 4.3.3. Smooth trend (loess) or straight lines are fitted to the data points to highlight the trend.

4.4.1.1 Faralla et al. (2012)

Figure 4.16 shows the relationship between the sign effect size and amount ratio. The points are coloured by the sooner amount. Each panel refers to a unique combination of the sooner and later delay in days. There is no clear relationship between the sign effect size and amount ratio.

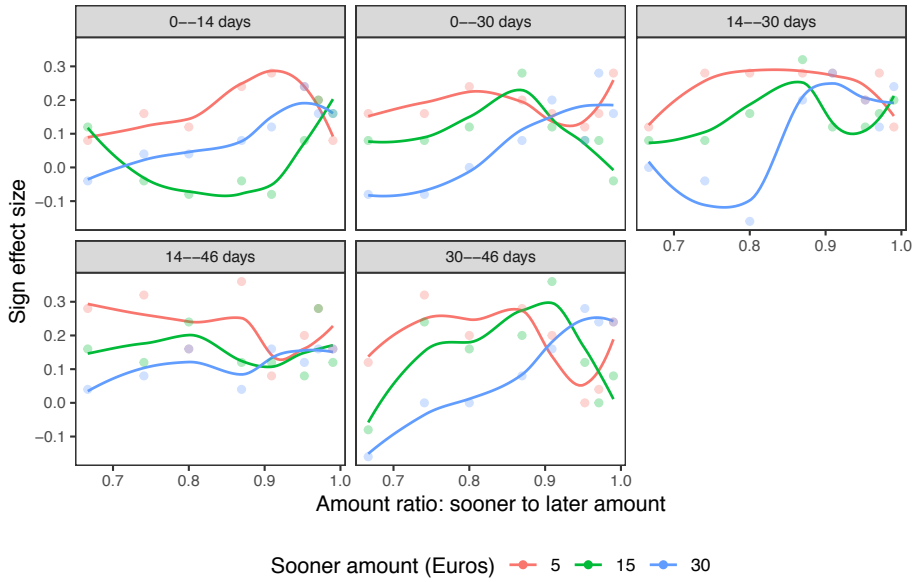


Figure 4.16: Sign effect size by amount ratio with smooth trend lines for Faralla10. The points are coloured by the sooner amount. Each panel refers to a unique combination of the sooner and later delay in days.

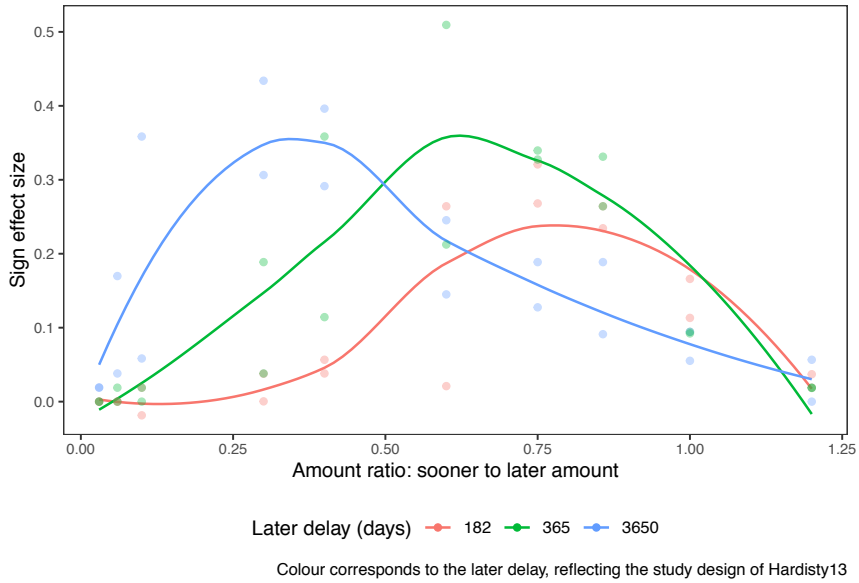


Figure 4.17: Sign effect size by amount ratio with smooth trend lines for Hardisty13. Points and lines are coloured by the later delays.

4.4.1.2 Hardisty et al. (2013)

Figure 4.17 shows the relationship between the sign effect size and amount ratio. Points and lines are coloured by the later delays. There is no clear relationship between the sign effect size and the amount ratio.

4.4.1.3 Han and Takahashi (2012)

Figure 4.18 shows the relationship between the sign effect size and amount ratio. Each panel represents a time delay. The points and smooth lines are coloured by the orrder in which the sooner amounts were presented. There relationship appears to be positive up to 30 days, before becoming more curvilinear up to 1825 days (5 years). There is no discernable relationship for the longest delay, i.e. 25 years. There appears to be systematic differences by the presentation order.

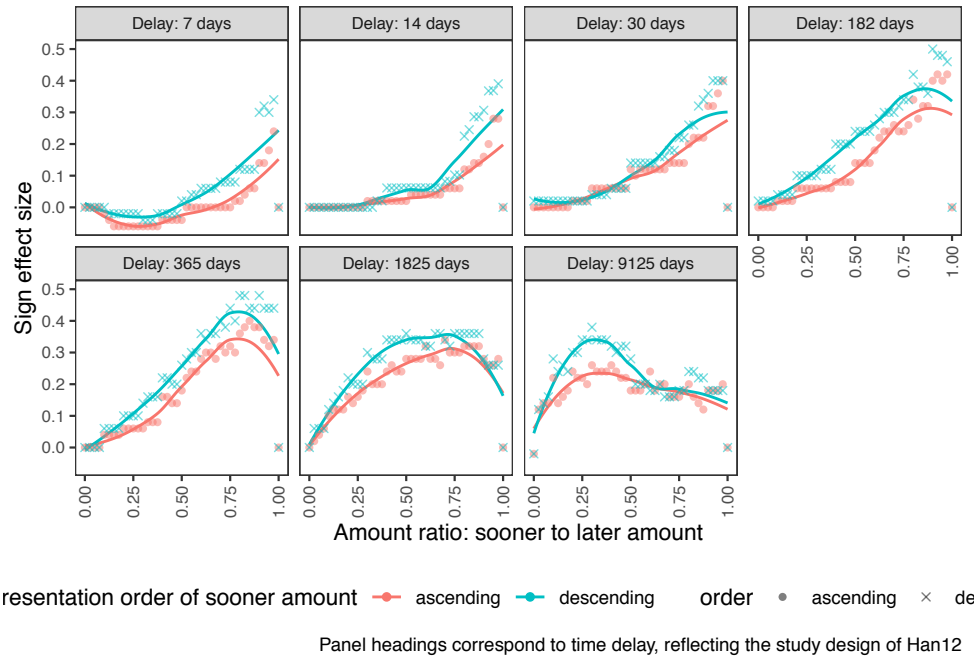


Figure 4.18: Sign effect size by amount ratio with smooth trend lines for Han12.

4.4.1.4 Xu et al. (2009)

Figure 4.18 shows the relationship between the sign effect size and amount ratio. Each panel represnts a unique combination of the sooner and later delay in days. A straight line is fitted to the data points in each panel.

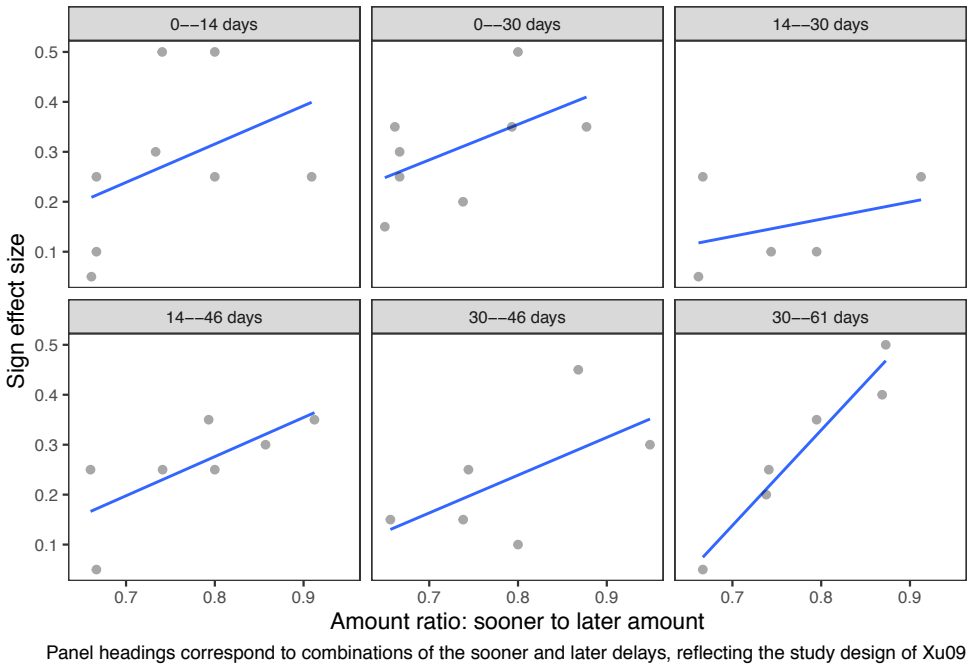


Figure 4.19: Sign effect size by amount ratio with straight lines for Xu09.

4.4.2 Individual-level: Exploratory data analysis

Individual participant data (IPD) are available for 6 out of 9 studies (see Table 4.2).

4.4.2.1 Indifference points

Calculating an indifference point for each individual is central to estimating discount rates. Of the 6 studies with IPD, 2 studies provided indifference points, which is a summary measure, for each participant. To calculate each choice each participant made, an assumption was made that each participant had a single swithing point, and thus one unique indifference point for each factor in the study design.

The remaining 4 studies provided raw choices of each participant, i.e. each choice each participant made for all questions. It is worth understanding the ease of obtaining accurate indifference points for each subject as these indifference points are used to calculate discount rates. This section uses 2 datasets for an empirical investigation into the issues raised in Section 3.2.3.3, regarding the challenges of obtaining accurate indifference points.

4.4.2.1.1 Faralla et al. (2012) Faralla et al. (2012) has a 3 (sooner amount) $\times$ 5 (sooner and later delay combination) $\times$ 2 (sign) study design. Each participant answered 240 questions in total, half of which were gains and the other half losses. This meant that each participant answered 8 (i.e. $\frac{240}{3\times5\times2}$) questions in a 'trial' resulting from a unique combination of the study design factors. For each participant, an indifference point can be calculated in each of the 30 trials. To calculate an indifference point, a participant must have only one 'switch point'. This means that a participant can only switch his/her preference for the sooner or later option once on a trial of 8 questions.

Figure 4.20 shows each of the 8 choices each participant made when the sooner amount was 5 euros available today and the later amounts were available in 2 weeks. The panel headings provide information on the gender, education status (ungraduate or post graduate) and study identification number. The participants are ordered by increasing number of switch points.

Some observations worth highlighting in Figure 4.20. Most participants do not have a single switching point. It would be difficult to estimate an accurate discount rate. Within-participant choices are heterogeneous. For example, 'M_UG_econ_02' does not switch for losses but switches several times for gains (see panel: row 4, column 4). There is also significant heterogeneity of choices made across participants.

Figure 4.21 shows a cumulative density plot of the number of switch points for each participant in each 'trial', i.e. a unique set of 8 questions based on the factorial study design. There are a total of 30 trials. Participants are ordered by the highest proportion of a single switch point.

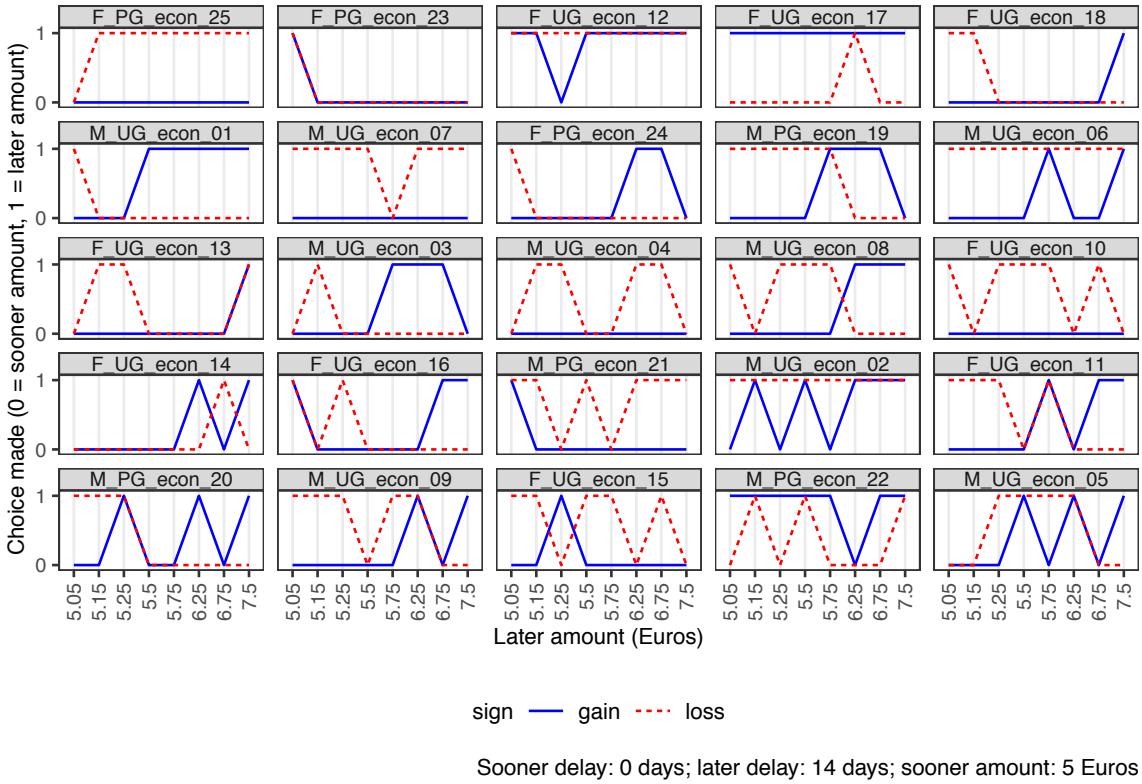


Figure 4.20: Choices each subject made when the sooner amount of 5 Euros was available today and the later amounts were available in 2 weeks.

Table 4.3: Percentage of switch points for gains and losses across all participants and trials.

	Number and percentage of switch points						
sign	0	1	2	3	4	5	6
gain	18.4	36.8	20.0	15.7	4.8	3.7	0.5
loss	17.1	29.3	17.1	20.8	10.4	4.8	0.5

Table 4.3 shows the proportion of switch points for gains and losses across participants and trials rounded to two significant figures. Since there were eight questions in a trial, the maximum switching points could be seven. There was no switching of preferences about 18% of the time. A single switch point occurred only 37% of the time for gains and 29% for losses. This represents how often an indifference point

can be calculated. There was a higher number of switching points in losses than gains, indicating greater inconsistency in preferences.

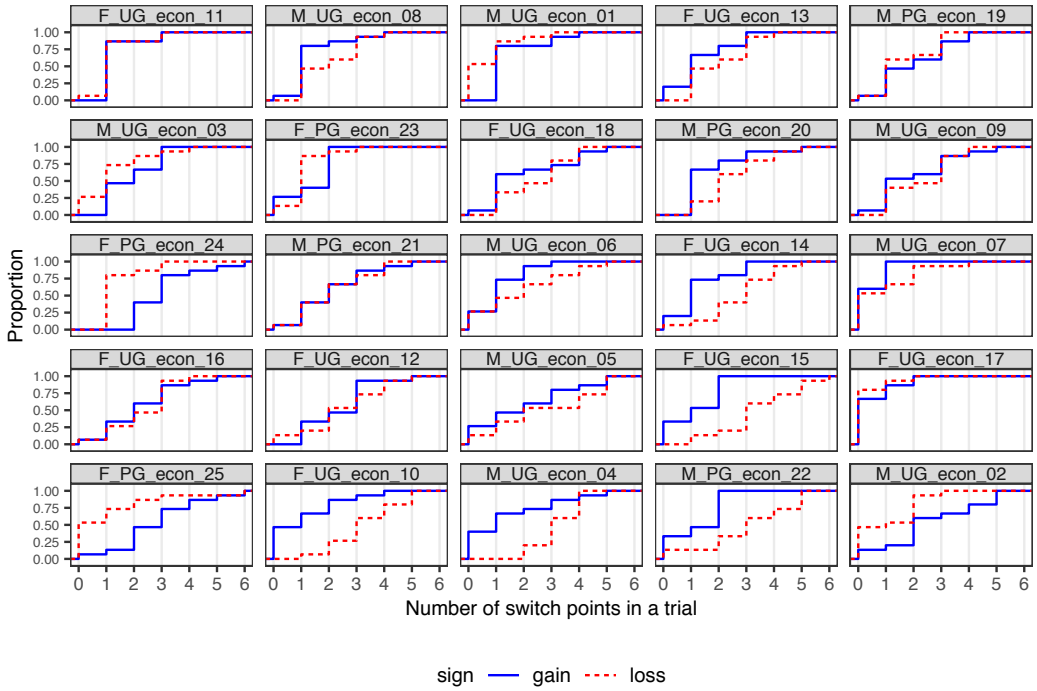


Figure 4.21: Cumulative density of the number of switch points across all 30 trials.

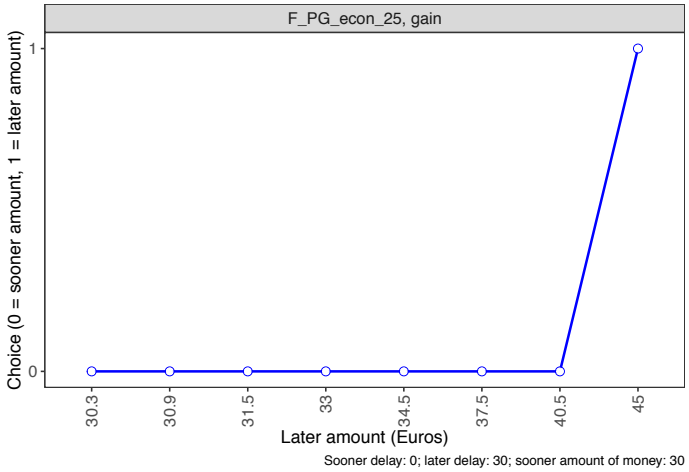


Figure 4.22: Illustration of a last minute switch.

There were 248 occurences of a single switch point in total (across gains and losses). Of these 248 occurences, 39 (15.7%) were 'last minute switches'. A last minute

switch indicates a single switching point on the last question in a trial. In a trial of 8 questions, the sequence of choices could either be: '0, 0, 0, 0, 0, 0, 0, 1' or '1, 1, 1, 1, 1, 1, 1, 0'. An example of the former sequence is shown in Figure 4.22.

4.4.2.1.2 Xu et al. (2009) Table 4.4 shows the number of question pairs each participant saw for each unique combination of the sooner and later time delay, which is based on the study design. Within each sooner and later delay combination, participants were presented with different sooner and later money amounts.

Table 4.4: Number of question pairs presented to each participant for each unique combination of factors based on the study design.

	ID	Sooner delay	Later delay	No. question pairs per participant
1	Xu09	0	14	8
2	Xu09	0	30	8
3	Xu09	14	30	5
4	Xu09	14	46	7
5	Xu09	30	46	6
6	Xu09	30	61	6

It is possible to calculate an indifference point for a participant in each unique combination of the sooner and later delay. Although participants were presented with different amounts of money to choose from, the difference between the sooner and later money amounts can be used. This is analogous to a 'titration' set up where differences in money amounts get closer.

Figure 4.23 shows the choices each participant made for a unique combination where the sooner amount was available immediately and the later amount in 2 weeks. The panel headings provide information on the gender and study identification number. The participants are ordered by increasing number of switch points.

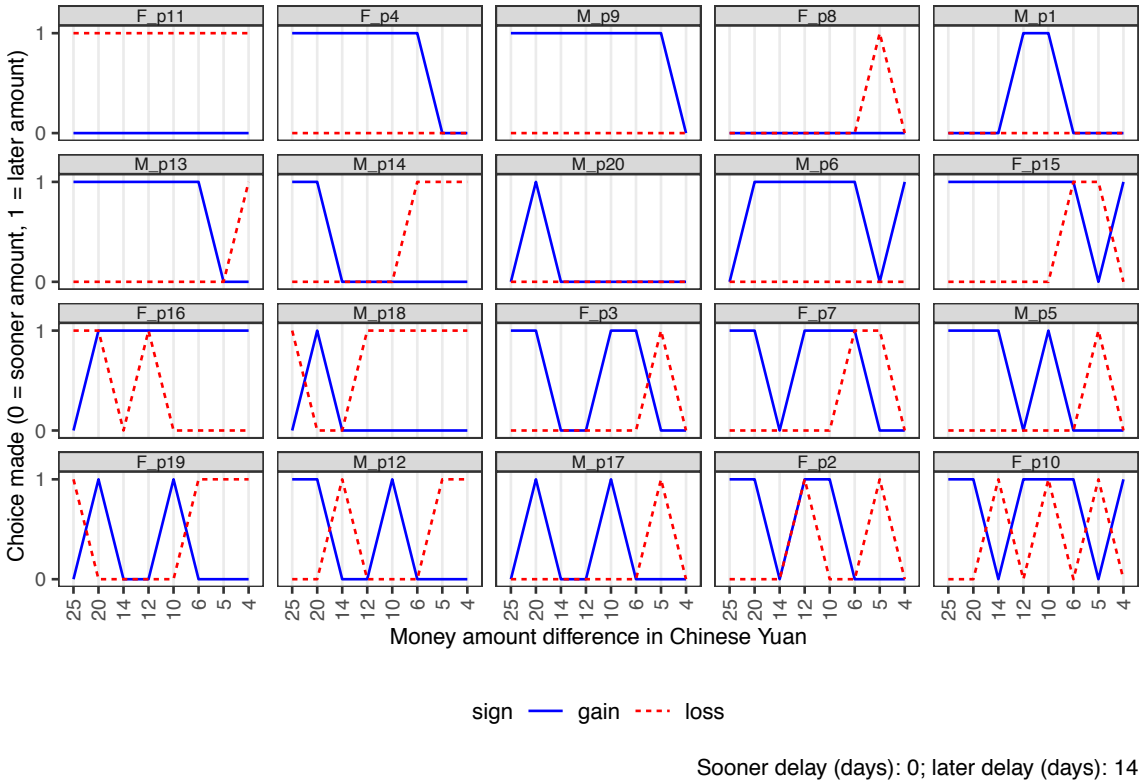


Figure 4.23: Choices for each participant when the sooner amount was available immediately and the later amount in two weeks. The amount difference was calculated by subtracting the sooner amount from the later amount.

Figure 4.24 shows a cumulative density plot of the number of switch points for each participant in each 'trial', i.e. a unique set of questions based on the factorial study design. There are a total of 12 trials. Participants are ordered by the highest proportion of a single switch point.

Table 4.5 shows the proportion of switch points for gains and losses across participants and trials rounded to two significant figures. The number of questions in a trial varied from 5 to 8 depending on the combination of sooner and later time delays (see Table 4.4). There was no switching of preferences 17% of the time for gains and 33% for losses. A single switch point occurred only about 30% of the time. This represents how often an indifference point can be calculated accurately. There was a higher number of switching points in gains than losses, indicating greater inconsistency in preferences.

Table 4.5: Percentage of switch points for gains and losses across all participants and trials.

	Number and percentage of switch points						
sign	0	1	2	3	4	5	6
gain	16.7	30.8	19.2	23.3	6.7	1.7	1.7
loss	33.3	29.2	21.7	11.7	3.3	0.0	0.8

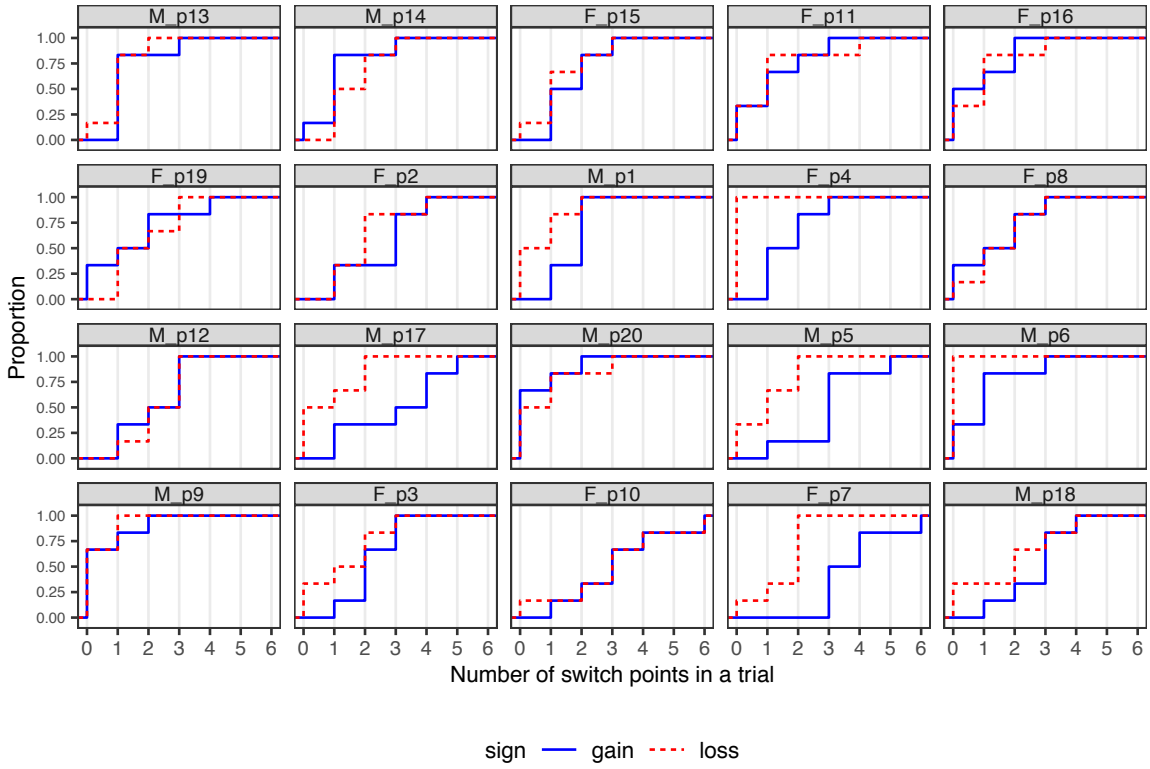


Figure 4.24: Cumulative density of the number of switch points across all 12 trials.

There were 72 occurences of a single switch point in total (across gains and losses). Of these 72 occurences, 7 (9.7%) were 'last minute switches'. A last minute switch indicates a single switching point on the last question in a trial. In a trial of 8 questions, the sequence of choiceLL could either be: '0, 0, 0, 0, 0, 0, 0, 1' or '1, 1, 1, 1, 1, 1, 1, 0'. An example was shown in Figure 4.22.

4.4.2.1.3 Discussion This section provided an empirical investigation into the issues of obtaining accurate indifference points for each participant, which were discussed in Section 3.2.3.3. To calculate an accurate discount rate, a participant should only have one switch point in each trial. This would be reflected in Figures 4.21 and 4.24 as a single vertical bar from 0.00 to 1.00 when the number of switch points is 1 on the x-axis.

However, in both figures, there are no participants with this profile. Instead, participants tend to have multiple switching points and very heterogeneous profiles. From Tables 4.5 and 4.3, only about a third of trials have a single switching point. Of these trials with a single switching point, 39 (15.7%) in Faralla et al. (2012) and 7 (9.7%) in Xu et al. (2009) were 'last minute' switches. It is worth asking if we should assume that subjects who switch preferences in the last question of a trial will remain consistent with their choices over more questions, given the inconsistency demonstrated.

4.4.2.2 Estimating the sign effect

One way of estimating the sign effect is to model individual choices on each question pair. In all studies, the questions for gains are identical to questions for losses, except for the sign of the amounts offered. Thus, it is appropriate to model individual responses to question pairs.

We attempt to model the sign effect from a discounting approach, which was discussed in Section 3.2.1.2. Extending that section, let $S^{+}_{i,j}$ be an indicator function, which takes the value 1 if the sooner gain is chosen and 0 if the later gain is chosen for participant i on question pair j. Let $L^{-}_{i,j}$ be an indicator function, which takes the value 1 if the later loss is chosen and 0 if the sooner loss is chosen for participant i on question pair j. Then each participant's response to a given question pair can be represented in Table 4.6 (reproduced from Section 3.2.1.2).

Table 4.6: Modelling individual participant responses to a question pair using a discounting approach.

Discounted?		$S^+_{i,j} - L^-_{i,j}$	Discounting interpretation
Gain	Loss		
Yes	Yes	1 - 1 = 0	Gain and loss discounted
Yes	No	1 - 0 = 1	Gain discounted only
No	Yes	0 - 1 = -1	Loss discounted only
No	No	0 - 0 = 0	No discounting

The sign effect has been described as gains being discounted more than losses. If one wanted to model the sign effect for each individual on each question pair, this description would translate to:

$$\text{Sign effect}_{i,j} = [P(S^+_{i,j} = 1) - P(L^-_{i,j} = 0)] > [P(S^+_{i,j} = 0) - P(L^-_{i,j} = 1)] \quad (4.2)$$

4.4.2.2.1 Discounting gains and losses: independent events? Xu et al. (2009) reported using a 'paired t-test on the percentages of subjects' choices of smaller/sooner options' for gains and losses. Hardisty and Weber (2009) reported using a t-test on the discount rates for gains and losses. In these cases, responses to gain questions are treated separately from responses to loss questions. Table 4.7 below shows the proportion of choices for sooner gains and later losses separately, and the estimated sign effect size, which is a difference between the two proportions. The table is arranged by studies with increasing effect size.

If the analysis were based on discounting on question pairs, rather than separate questions by sign, then responses to a question pair can be classified into 4 categories (see Table 4.6): gain discounting only, loss discounting only, both gain and loss discounting, and no discounting. Table 4.8 shows the proportion of responses to

Table 4.7: Proportion of sooner gains and later losses with the sign effect size as a difference in proportions. Proportions are to two decimal places.

Study	No. participants	Total question pairs	Sooner gain	Later loss	Sign es
Hardisty13	106	3,180	0.50	0.36	0.14
Faralla10	25	3,000	0.62	0.48	0.14
Han12	50	28,659	0.28	0.13	0.15
Xu09	20	800	0.50	0.24	0.26
Hardisty09_exp1	65	650	0.54	0.24	0.30
Hardisty09_exp2	118	1,180	0.56	0.23	0.33

question pairs classified into the 4 categories for each paper, the sign effect, and the proportion of question pairs used to calculate the sign effect.

Adding the proportion of 'disc: G and L' (both gain and loss discounting) to 'disc: Gain' (gain discounting only) and 'disc: Loss' (loss discounting only) would correspond to the proportion of sooner gain and later loss choices in the previous table (with some rounding error). The sign effect size is the same in both tables. However, this highlights the significant proportion of overlapping choices that discount *both* gains and losses on question pairs. On average, only 30% of questions pairs can be used to calculate the sign effect. This means that 70% of question pairs provide no information on the sign effect.

4.4.2.2.2 Discounting behaviour for each participant Figure 4.25 shows density plots of the four different discounting behaviours within each of the 384 participants. Discounting of losses peak between 0 and 0.25, suggesting that most participants display little discounting of losses. Discounting of gains and discounting of both gains and losses have similar densities with positive skews. The density for no discounting has a negative skew, indicating that this was the most common discounting behaviour for a participant.

Table 4.8: Summary of the sign effect and discounting behaviour between studies based on IPD. Percentages are rounded to the nearest whole number. The sign effect size (es) is a difference in proportions and is rounded to two decimal places. The table is arranged by increasing sign es. Sign obs refers to the percentage of observations with the sign effect.

		Percentage (%)					
Study	Total question pairs	disc: Gain	disc: Loss	disc: G and L	disc: None	Sign obs	Sign es
Hardisty13	3,180	15	1	35	49	16	0.14
Faralla10	3,000	24	10	37	28	35	0.14
Han12	28,659	17	2	11	69	19	0.15
Xu09	800	31	5	19	44	37	0.26
Hardisty09_exp1	650	31	1	23	46	31	0.30
Hardisty09_exp2	1,180	34	1	22	43	35	0.33

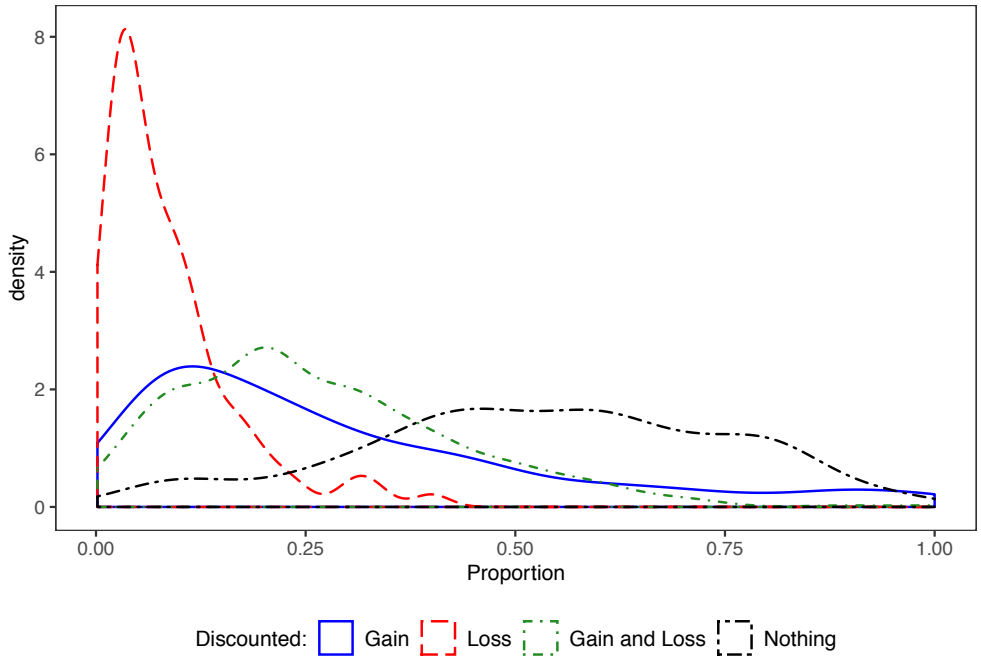


Figure 4.25: Within-subject variation in discounting behaviour across all studies.

Figure 4.26 extends Figure 4.25 by plotting the densities of the within-participant discounting behaviours for each of the 6 studies. There is significant heterogeneity in discounting behaviours within and between studies. For example, the density for no

discounting across studies can be positively or negatively skewed, or rather uniformly distributed.

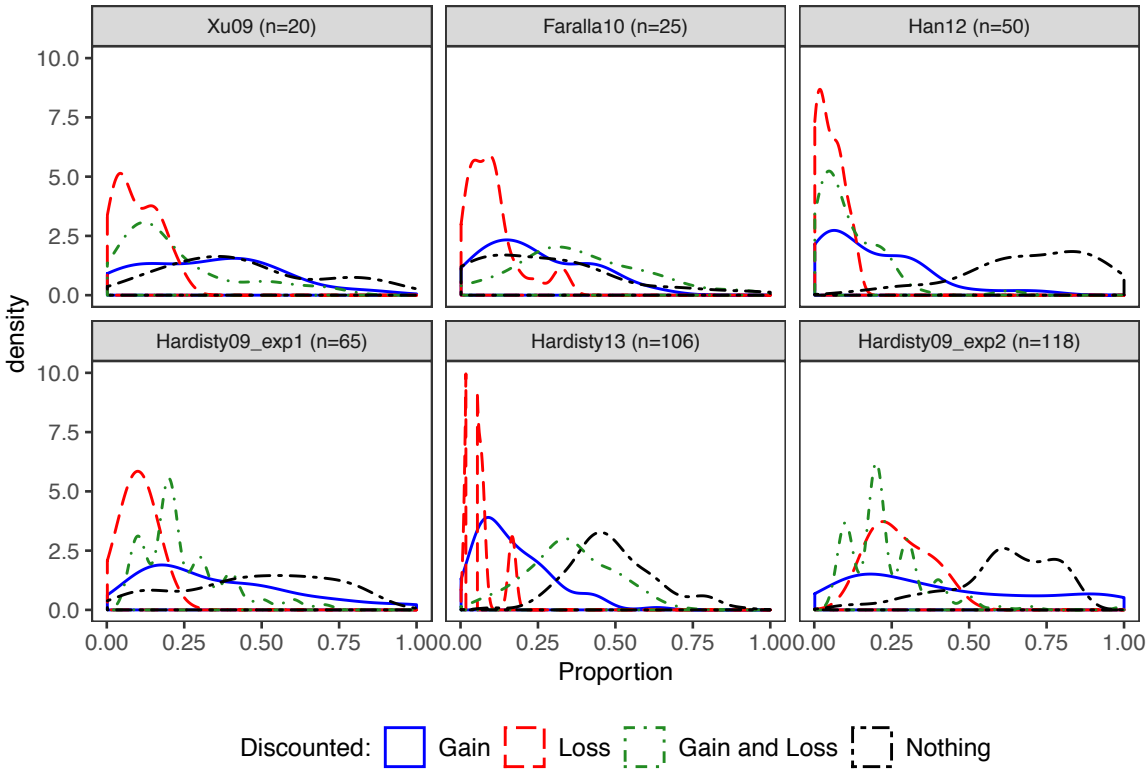


Figure 4.26: Density plots of within-subject variation in discounting behaviour for each study.

4.4.2.2.3 Informative responses Some scholars have suggested that the sign effect can be tested with a McNemar test, which involves comparing the off diagonals (cells 'b' and 'c') in the Table 4.9, where cell b refers to choosing the sooner gain and later loss (discount gain only), and cell c refers to choosing the later gain and sooner loss (discount loss only).

One assumption of the McNemar test is that there must have enough numbers in cells b and c, e.g. at least 5 in each cell. The cumulative density plot below shows participant responses to question pairs by discounting type in each paper. The number of participant responses is on the x-axis. In each plot, there is a vertical line on the x-axis that corresponds to a value of 5. For the coloured lines, 'Gain' refers

to cell b and 'Loss' to cell c.

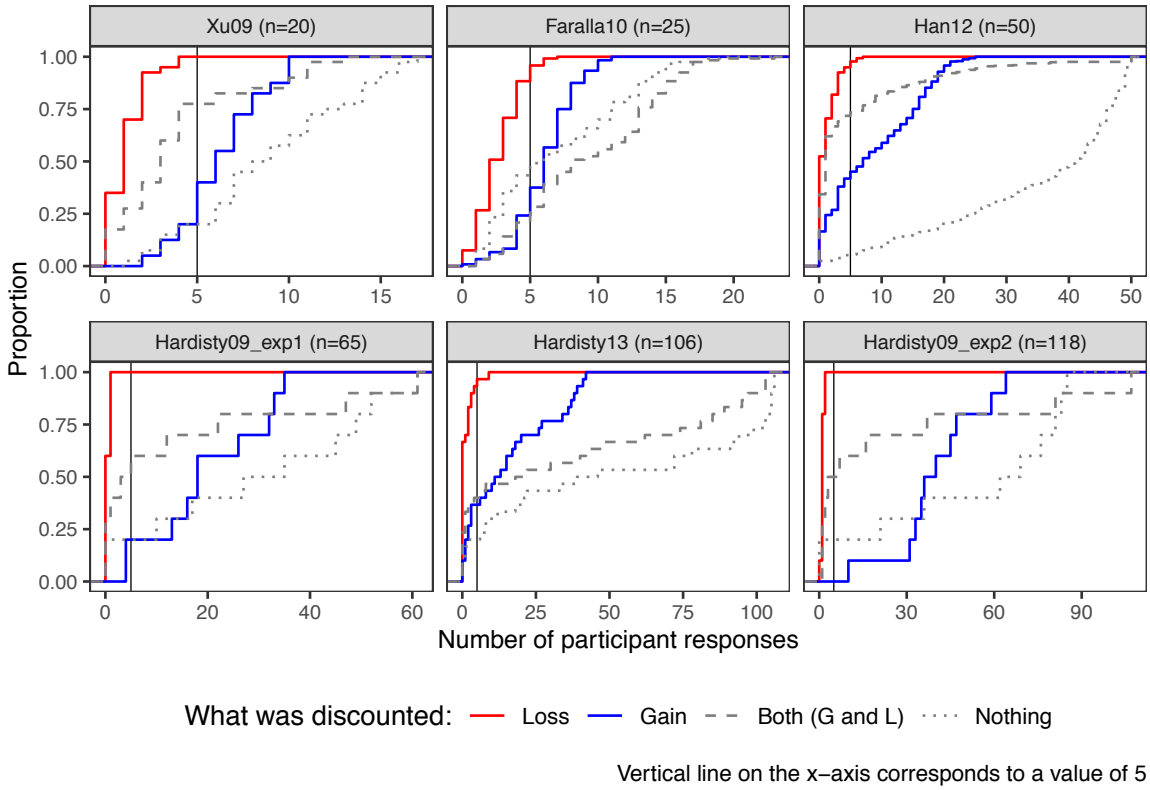


Figure 4.27: Cumulative density of participant responses on question pairs, coloured by type of discounting behaviour

Table 4.9: Illustration of responses to a question pair in a 2 x 2 table.

	Soon (-)	Late (-)
Soon (+)	a	b
Late (+)	c	d

The McNemar test might not be the most suitable statistical test to use. Across the studies, 98% of the 784 question pairs have 5 or fewer responses to either cells b or c. Further, 48% of the 784 observations have zero responses to either cells b or c, i.e. the cells are empty. Thus, almost all the data would be discarded if a McNemar test were used.

Also, the McNemar test is usually two-tailed, which would not be appropriate since the sign effect definition has only one direction. One possible alternative to the McNemar test is an exact binomial test. Because the sign effect has been defined as the discount rate for gains being *greater* than the discount rate for losses, the null and alternative hypotheses can be set up as:

$$H_0 : P_b = P_c \tag{4.3}$$

$$H_1 : P_b > P_c \tag{4.4}$$

However, 74 out of 784 (9%) question pairs must be dropped as both cells b and c are empty. The 74 empty observations were from 2 studies, with 96% from Han12, and the rest from Hardisty13.

If, for illustration purposes, an exact binomial test were to be performed on each of the remaining 710 question pairs, then 56% of questions pairs would have a statistically significant result ($p < 0.05$), which reduces to 13% after applying the Bonferroni correction. This assumes independent observations, which is very unlikely as the same participants answer multiple question pairs in a paper.

The sign effect could be modelled at the individual level. However, the data suggest some challenges. For example, 83% of the 384 participants from the 6 studies do not discount gains or losses on any question pair, i.e. empty cells for either gain or loss discounting. Table 4.10 below shows this breakdown by study.

Figure 4.28 shows a jittered strip chart of the sign effect size for each participant in each study. Each point represents one participant's sign effect size and there are 384 points. The different shape and colour of the points indicate if a participant had zero discounting for gains or losses on all question pairs (red crosses) or not (black circles). The dotted blue horizontal line represents the zero line of no effect. Studies are arranged by increasing median value.

Table 4.10: Breakdown of number and percentage of participants who did not discount any gain or any loss on any question pair in each study.

Study	Participants who did not discount any gain or any loss (%)	Total no. participants
Faralla10	0.0	25
Xu09	50.0	20
Han12	58.0	50
Hardisty13	90.6	106
Hardisty09_exp1	100.0	65
Hardisty09_exp2	100.0	118

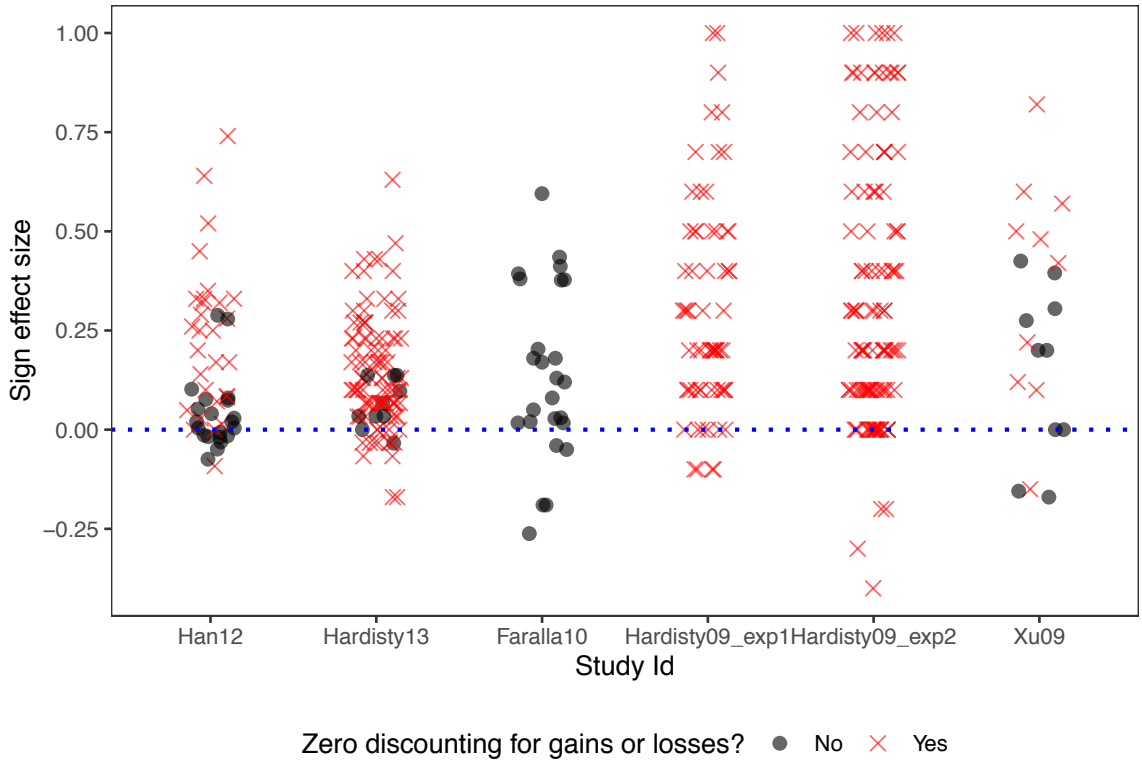


Figure 4.28: Sign effect size for each participant coloured by whether the participant had zero discounting for gains or zero discounting for losses across question pairs. Each point is jittered horizontally.

Of the 384 participants, 318 (83%) participants had zero discounting for gains or zero discounting for losses across the question pairs they responded to, as most

(77%) participants did not discount any loss. Participants with zero discounting for gains or losses tended to display greater sign effect sizes, suggesting that these participants tended to only discount gains and not losses. The median sign effect size for participants who had zero discounting for gains or losses was 0.2 (range: -0.4 to 1) compared to a median value of 0.045 (range: -0.3 to 0.6) for the 66 participants who discounted gains and discounted losses.

4.4.3 Formal statistical modelling

Based on the exploratory data analysis of the sign effect using IPD in Section 4.4.2.2, there are insufficient data to appropriately model the sign effect. Hence, formal statistical modelling of the sign effect will not be conducted. An appropriate analysis of the sign effect in our sample requires IPD as aggregate data do not provide information about the joint probability of discounting gains and losses. There were IPD for 6 studies, with a total of 37,469 responses to 784 question pairs from 384 participants.

IPD can be used to analyse the sign effect on each unique question pair (question-level analysis). On each question pair, participants' responses can belong to one of four mutually exclusive categories: discount gain, discount loss, discount gain and loss, and no discounting. If represented in a 2×2 contingency table, discount gain would represent cell 'b' and discount loss cell 'c' in the off-diagonal. Of the 784 unique question pairs, 98% have 5 or fewer responses in cells b or c, and 48% had zero responses in either cell.

IPD can also be used to model each choice each individual makes on each question pair. Of the 384 participants, 83% did not discount any gains or losses on any question pair. Thus, only 17% of participants can be included in any formal analysis.

4.5 Discussion

According to influential narrative reviews, the sign effect is a well-established "anomaly" in intertemporal choices. 'Many studies have concluded that gains are discounted at a higher rate than losses' and 'this pattern is prevalent in the literature' (Read 2004; Frederick, Loewenstein, and O'Donoghue 2002, 362–63). The sign effect is considered to be a 'relatively robust effect', although some studies have reported greater discounting for losses than gains (Read 2004).

This chapter aimed to conduct a quantitative systematic review of the sign effect. It focussed specifically on studies that presented participants with binary choice question pairs involving monetary gains and losses. A question pair is a pair of questions that is identical in all respects except the sign, i.e. the delays are identical and the amounts are identical except in one question the amounts are gains and in another question they are losses.

Individual participant data (IPD) were available for 6 studies from 6 papers. The data consisted of 37,469 responses to 784 unique question pairs from 384 participants. That is, IPD were available for a total of 74,938 responses.

Results from this quantitative systematic review suggest that existing evidence for the sign effect is based on inappropriate statistical analysis. Studies tend to analyse gains and losses separately, e.g. taking the difference in proportions for gains and losses, or calculating a discount rate for gains and losses separately. When the difference in proportions is analysed this way across all studies, a relatively strong sign effect is observed. However, this is based on the inappropriate assumption that discounting for gains and discounting for losses are two independent events.

If responses to each question pair were analysed, then the median proportion that discounting is observed in both the gain and the loss question is 0.22 (range: 0.11 to 0.37), providing evidence for a sign effect. When the sign effect is analysed with IPD at the question level, the median effect size across the 6 studies is 0.2 (range: 0.14 to 0.33), providing evidence for a sign effect. When the sign effect is calculated by subtracting instances where a participant discounted both gains and losses on a

question pair, then the median proportion of question pairs where the sign effect is observed is 0.33 (range: 0.16 to 0.37). This means only about a third of question pairs can be used to test for the sign effect.

The median proportion of question pairs where only the gain is discounted is 0.27 (range: 0.15 to 0.34). The median proportion of question pairs where only the loss is discounted is 0.016 (range: 0.0062 to 0.1). Finally, the median proportion of question pairs where nothing is discounted is 0.45 (range: 0.28 to 0.69).

When the responses are analysed on the individual level, then within individuals the median proportion of discounting both the gain and loss on a question pair is 0.2 (range: 0 to 1). The median proportion of discounting the gain only on a question pair is 0.2 (range: 0 to 1). The median proportion of discounting the loss only on a question pair is 0 (range: 0 to 0.4). Of the 384 participants, 83% did not discount any gains or losses on any question pair. The median proportion of no discounting on a question pair is 0.5 (range: 0 to 0.98). The median effect size is 0.17 (range: -0.4 to 1). However, of the 384 participants, 318 (83%) participants had zero discounting for gains or zero discounting for losses in all the question pairs they responded to. These participants tended to have a much higher effect size than participants who discounted gains and losses (median: 0.2 vs. 0.05).

This chapter makes several contributions to the existing literature on the sign effect. It provides a definition based on individual responses to question pairs, which allows for a much more nuanced description of discounting behaviour. The literature has focussed on the average discounting of gains and losses, while overlooking occurences when both gains and losses are discounted, and when nothing is discounted. This chapter provides strong evidence that discounting of gains and discounting of losses are not independent events, which are assumptions underpinning the studies in our sample.

This chapter also highlights significant heterogeneity within and between participants. It is not uncommon for a participant to display several types of discounting behaviour in their choices, e.g. to discount gains, losses, both, and nothing. Within a study, the behaviour between participants also varies. These types of variation

are not reported in the literature, which tends to report averages only. Given the substantial heterogeneity within- and between-participants, summarising a participant's responses into one value, e.g. an average or a discount rate, might not be appropriate.

Results from this chapter reveals some rather surprising findings. The most common behaviour at both the individual and question level is that of no discounting. It was also uncommon for losses to be discounted at all. Reports of loss discounting are likely to have included the joint probability of gain and loss discounting.

This chapter also calls into question the convention of calculating indifference points for each participant. This requires participants to have a single switching point in their preferences. Based on raw IPD from two studies, no participant had a single switching point across all trials. There was significant heterogeneity within and between participants. A single switch point occurred only about a third of the time.

In behavioural science, there is a theory of 'trembling hands' where there will be some amount of 'random error' in participants' responses such that a participant's preferences may appear inconsistent within a trial when their true preference is consistent (Birnbaum and Bahra 2012). If this were true, it would be reasonable to expect a single switch point on most trials. However, Section 4.4.2.1 provides evidence against an argument that any apparent inconsistency is due to 'trembling hands'. The number of switch points within a participant varies substantially within and between trials.

One limitation of this chapter is that the studies in this sample might not be representative of the population of sign effect studies with question pairs. Another limitation is the small sample of 6 studies. More data can be collected and analysed. However, it is still interesting and valid to compare the results from this quantitative review with how the sign effect has been reported in the original studies and influential narrative reviews. Finally, the quality of the reported studies could have been assessed using a checklist. However, the appropriate checklist to use in these behavioural science systematic reviews remain up for debate.

Chapter 5

Statistical modelling of individual participant responses to monetary gains and losses

This chapter will model individual participants' responses to binary questions involving monetary gains and losses. Individual participant data (IPD) from six different studies will be used. The aim of this chapter is to understand the factors that influence individual participants' choices in the presence of gains and losses.

Each section will be one study with IPD. Each section will explain the study design, explore the data, fit multilevel models, and diagnose the performance of the models. There will be a focus on exploring potential heterogeneity within and between participants, which has not been adequately explored in the literature. The literature has focussed on aggregate-level analyses, e.g. by comparing group averages.

The literature has mostly focussed on estimating a sign effect when there are question pairs of gains and losses. However, the previous chapters demonstrated the considerable difficulties in accurately defining, estimating and modelling the sign effect in the classical sense. This chapter will not test the sign effect in its classical sense but will estimate the extent to which the sign of choices, i.e. whether choices are presented as gains or losses, influence choices. This contributes to the goal of

understanding how intertemporal choices are made (Read 2004).

5.1 Faralla et al. (2012)

This data set has 6,000 rows and 25 participants. Each participant was asked 240 questions, half of which were gains and half losses. The study had an implicit within-participant factorial design: 2 (sign: gain, loss) × 3 (sooner amounts: 5, 15, 30) × 5 (sooner-later time delay combination: 28-42, 14-42, 14-28, 0-14, 0-28). There are a total of 30 factors (2 x 3 x 5). This means that there are 8 questions for each unique combination of the 30 factor levels. Theoretcally, responses to each unique combination of 8 questions can be used to caclulate an indifference point for a participant, giving 30 indifference points per participant (15 for gains and 15 for losses).

5.1.1 Exploratory data analysis

Of the 25 participants, 12 (48%) were women. There were 18 (72%) participants who were undergraduate students and 7 postgraduates. The median age of the participants was 26 years old (range: 21 to 31 years). Postgraduates tended to be slightly older than undergraduates (median age: 27 years vs. 25 years).

Of the 12 women, 9 (75%) were undergraduate students. Of the 13 men, 9 (69%) were undergraduate students. The median age of men was the same as women (26 years).

5.1.1.1 Proportion of later choices

Figure 5.1 below shows the proportion of later choices for each participant. Each point represents the overall proportion of later choices for one participant. The colour and shape of the points indicate whether the proportion is for gains (blue circle) or losses (red crosses). There is a dotted vertical line at 0.5 on the x-axis to highlight the points to the left and right of the line.

From Figure 5.1, two-thirds (66%) of the proportions are smaller than 0.5, which implies that most choices are for the sooner option. Only 6 points for gains are greater than 0.5 compared to 11 points for losses. The median proportion of later gains is lower than that of losses (0.37 vs. 0.46). Most (68%) points for losses are to the right of gains within participants. Men tend to choose the later loss more often than women (median: 0.53 vs. 0.38). For gains, men tend to choose the later option more than women (median: 0.42 vs. 0.31).

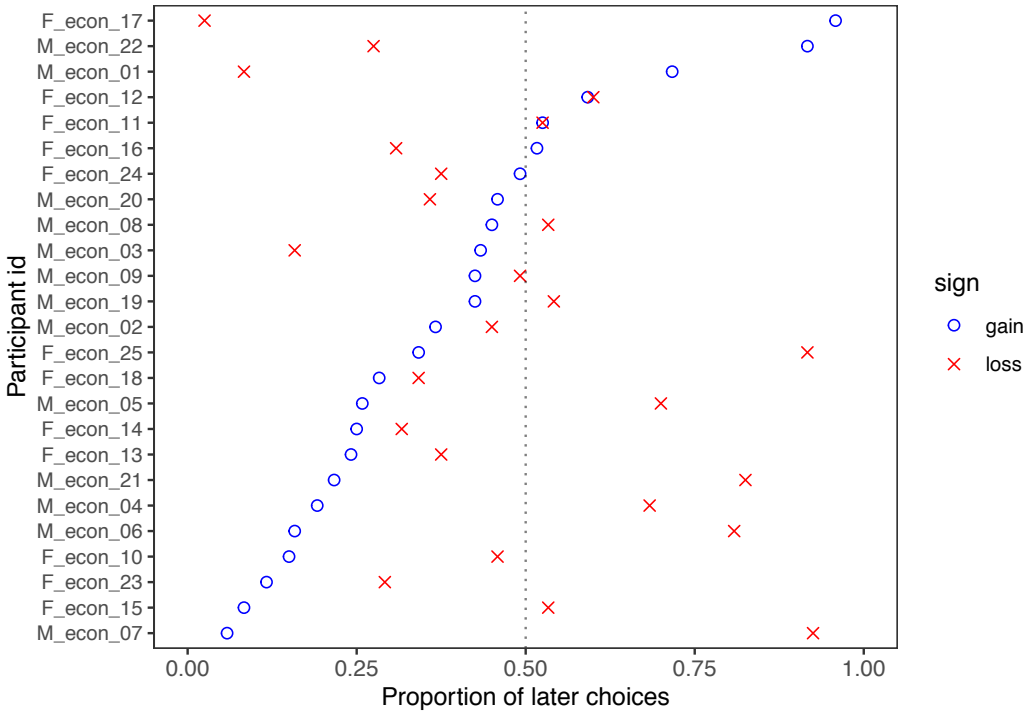


Figure 5.1: Strip chart showing the proportion of later choices for each participant. The participants on the x-axis are ordered by increasing proportion of later gains. The colour and shape of the points represent the two different values of sign.

Figure 5.2 shows density plots of the proportion of later choices coloured by sign for each of the 25 participants. The proportions were calculated based on the 30 proportions from each unique combination of 8 questions. There is substantial heterogeneity in the patterns of choices across participants. The amount of overlapping between the gain and loss density curves and the location of the densities vary substantially. For example, the top left and bottom right panels show participants who

do not have overlapping densities. However, the location of the densities for later gains and losses are at the opposite ends for the participants.

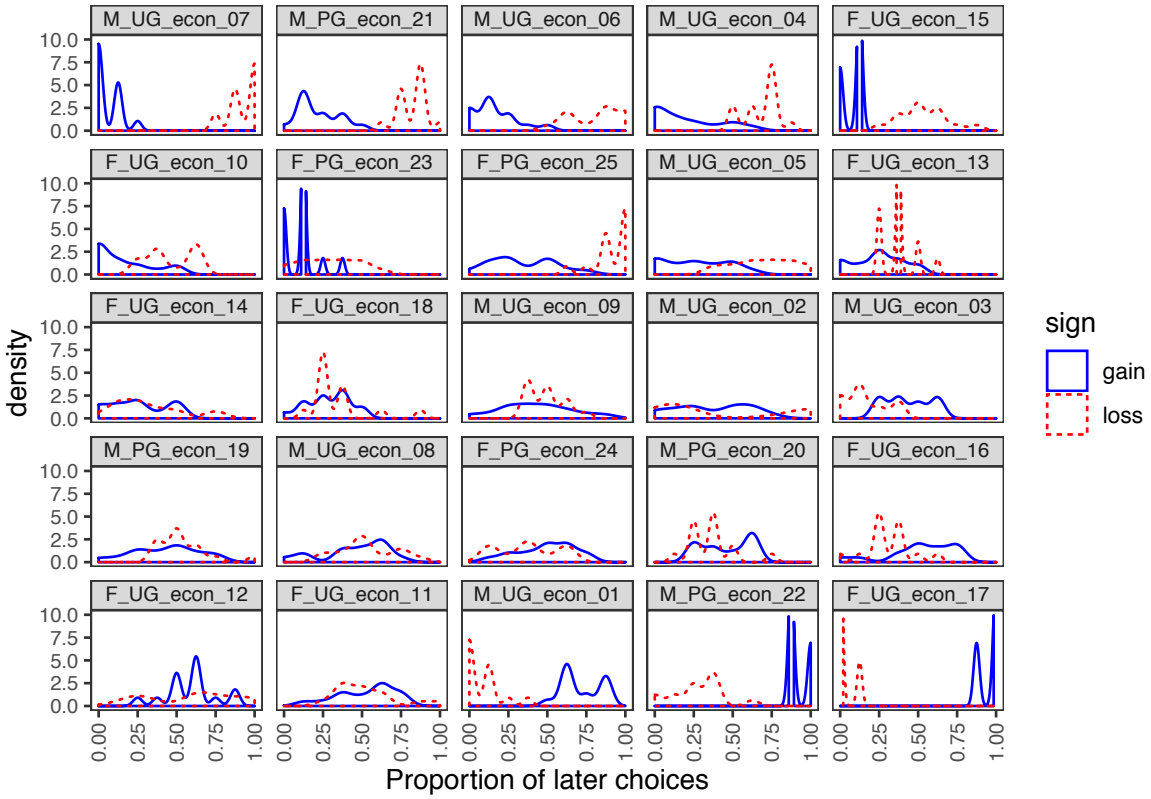


Figure 5.2: Density plots of the proportion of later choices, coloured by sign for each participant. The proportions are based on 30 proportions from each unique combination of 8 questions. The panels are arranged by increasing proportion choosing the later option for gains. Each panel heading provides information on the gender, university education status and study id for each participant.

5.1.1.2 Time delay

Figure 5.3 shows the relationship between choosing the later option and delay interval. The delay interval is calculated by taking the difference between the delays for the sooner and later amounts. There is an opposite relationship for gains and losses. For gains, most participants (21) tend to choose the later option less often with a longer interval. For losses, most participants (22) tend to chose the later option more often with a longer interval.

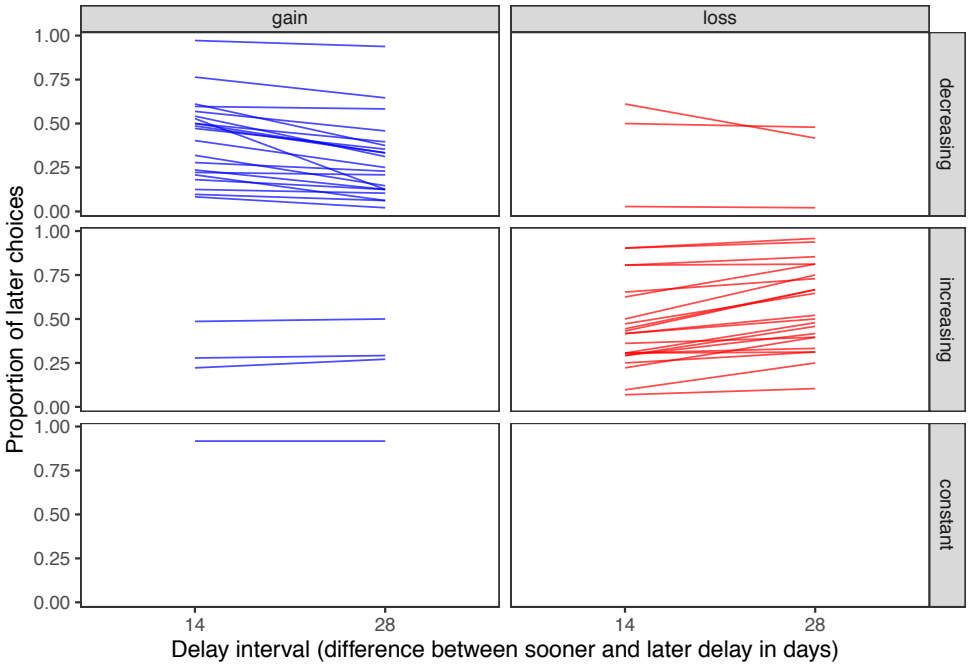


Figure 5.3: Line plots of the relationship between the proportion of later choices and the interval split by sign and whether the pattern was strictly decreasing, strictly increasing or constant over the interval values.

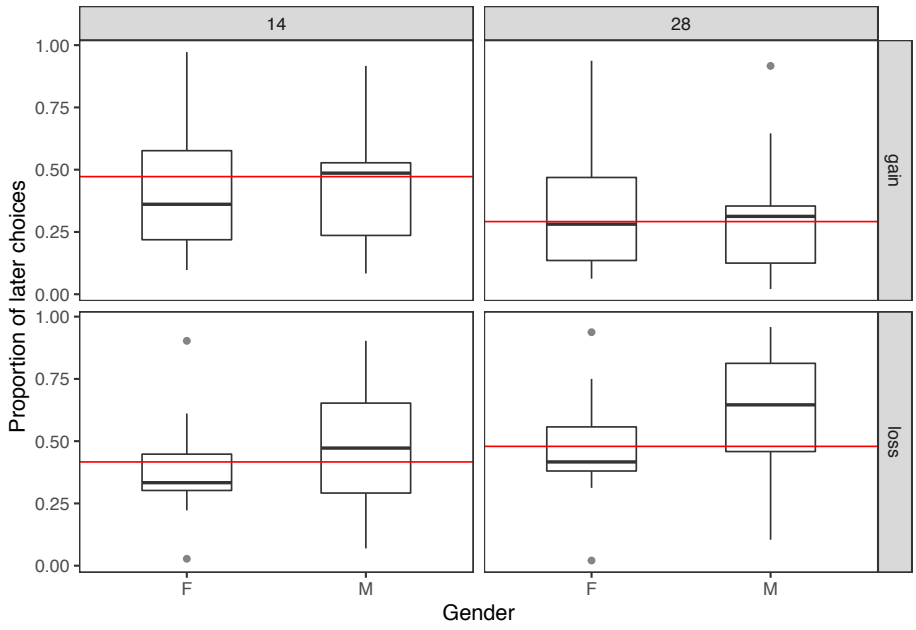


Figure 5.4: Boxplots of the proportion of later choices and gender by interval and sign. The horizontal red line represents the median value of that group (panel).

Figure 5.4 shows box plots of the proportion of later choices for each gender, which is on the x-axis, by sign and time interval. This helps visualise any potential interaction effects between these variables.

Overall, men have a slightly higher median value than women, indicating that men tend to choose the later choices slightly more often. For gains, the overall median proportion of later choices is higher for the shorter interval (14 days) compared to the interval of 28 days. There is an opposite relationship for losses: the overall median proportion of later losses tends to be lower for the shorter interval. For the shorter interval, the overall median line is slightly higher for gains than losses. For the longer interval, the overall median line is lower for gains than losses.

5.1.1.3 Money amounts

Figure 5.5 shows the relationship between the proportion of later choices per participant and the sooner amount of money split by sign and pattern of relationship. The pattern could by decreasing, i.e. the proportions are decrasing over the range of sooner amounts, increasing or mixed. The proportions are based on a partial factorial design (for the sooner amounts and sign only), meaning each proportion is calculated from a participant's responses to 40 questions.[1]

For gains, participants tend to choose the later option more often as the sooner amount increases. Participants display different patterns of choices for losses. Most participants tend to choose the sooner loss more often as the sooner amount increases although there are four participants who choose the later loss more often, which implies that these four participants tend to prefer to wait and pay a greater amount of money than pay a smaller amount of money sooner.

Figure 5.6 shows the relationship of later choices and sooner amount for each participant. There is substantial heterogeneity in choices for gains and losses across participants.

[1]There are 40 questions because: 240 questions for each participant divided by 2 (sign) x 3 (sooner amount) factors

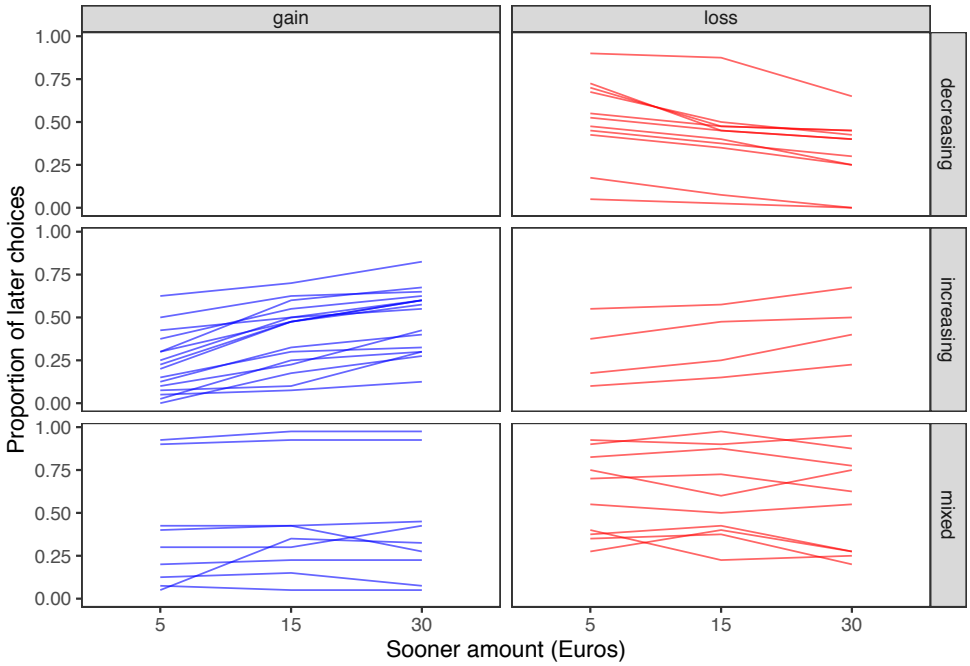


Figure 5.5: Line plots of the relationship between the proportion of later choices per participant and the sooner amount of money split by sign and whether the pattern was strictly decreasing, strictly increasing or mixed over the xs values.

Table 5.1: Number of questions for each unique value of amount ratio rounded to two decimal places.

Amount ratio (rounded to 2 d.p.)	Number of gain questions	Number of loss questions
0.67	15	15
0.74	15	15
0.8	15	15
0.87	15	15
0.91	15	15
0.95	15	15
0.97	15	15
0.99	15	15

The amount ratio is defined as the sooner amount divided by the later amount. Table 5.1 shows the number of questions for each unique amount ratio rounded to

two decimal places. The amount ratio is defined as the sooner amount divided by the later amount. It is rounded to get the proportion of later choices for the purposes of exploring the data.

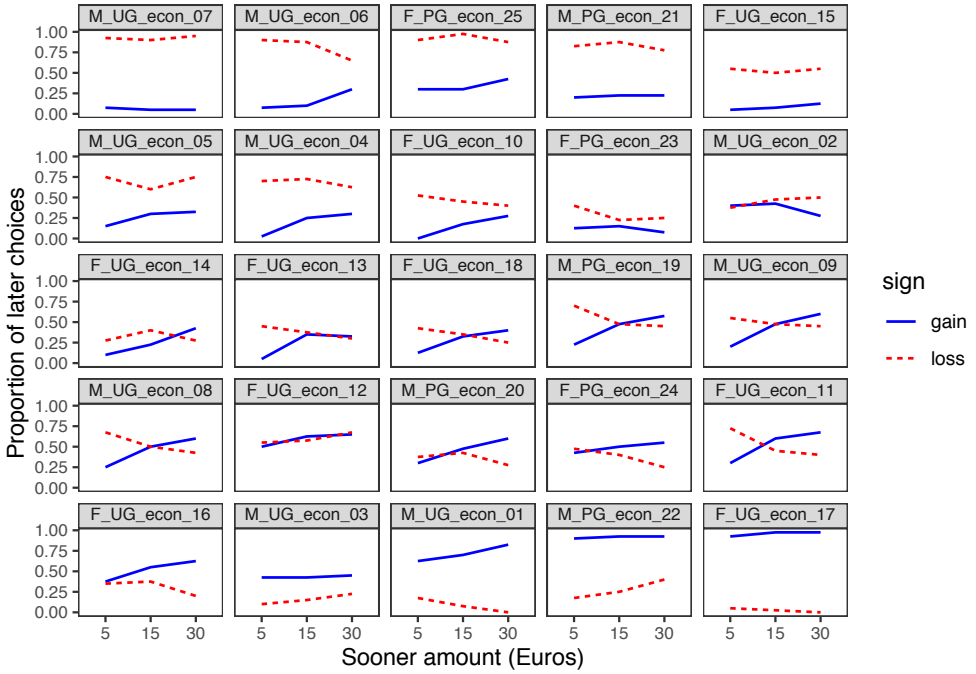


Figure 5.6: Line plots of the relationship between the proportion of later choices and sooner amount coloured by sign for each participant. Each proportion is based on a single summary value for the particular sooner amount. Each panel heading provides information on the gender, university education status and study id for each participant.

Table 5.2 shows the proportion of later choices aggregated across all participants for each unique value of amount ratio. As the amount ratio increases, the proportion of later choices decreases for gains but increases for losses.

Figure 5.7 shows the relationship between the proportion of later choices across all participants and amount ratio by the 3 different sooner amounts. The left panel shows the relationship for gains and the right panel for losses.

Table 5.2: Percentage of later choices across all participants for each unique value of amount ratio.

Amount ratio (rounded to 2 d.p.)	Percentage of later gains (%)	Percentage of later losses (%)
0.67	66	29
0.74	57	31
0.80	54	35
0.87	40	42
0.91	35	48
0.95	22	64
0.97	19	64
0.99	15	69

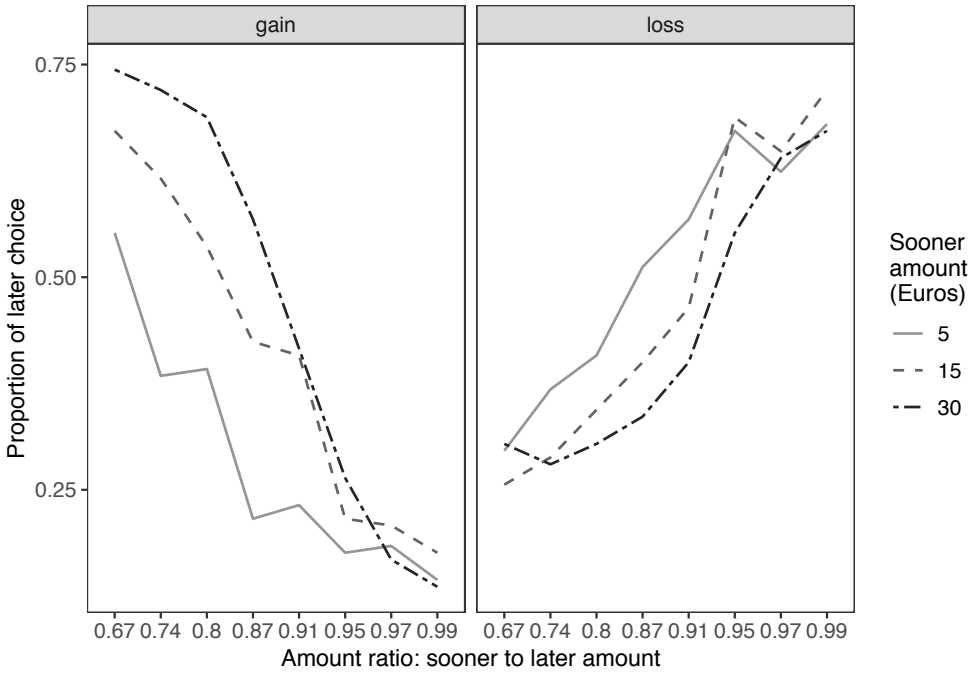


Figure 5.7: Line plots of the proportion of later choices and amount ratio by the 3 different sooner amounts and by sign.

The relationship between the proportion of later choices and amount ratio is negative for gains but positive for losses across the three different sooner amounts. The steepness of the lines varies slightly by the sooner amount. For example, the proportion

of later gains tends to be highest for the largest sooner amount (30 Euros), followed by the middle sooner amount (15 Euros) and then the smallest amount, with the opposite for losses, across the amount ratios.

Figure 5.8 shows the relationship between the proportion of later choices and amount ratio for each participant. Participants are ordered by increasing proportion of later gains. Participants tend to choose the sooner gain but later loss as the amount ratio increases. Some participants almost always choose the same option across amount ratios and the rate of change varies across the other participants.

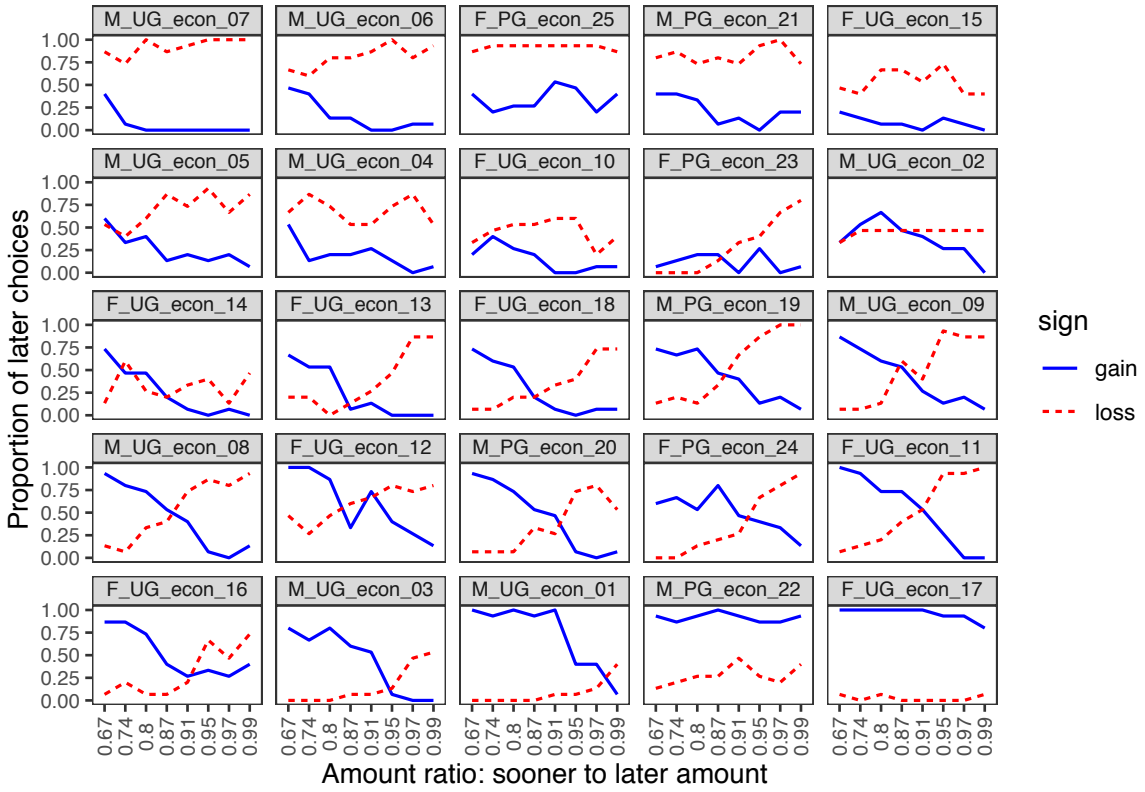


Figure 5.8: Line plots showing the proportion of later choices and amount ratio for each participant coloured by sign.

Figures 5.9 and 5.10 show the relationship between the proportion of later choices and amount ratio by the different sooner amounts for each participant. Figure 5.9 is for gains only (blue lines), while Figure 5.10 is for losses only (red lines). It is worth highlighting that the relationship by the different sooner amounts are heterogeneous

between and within participant.

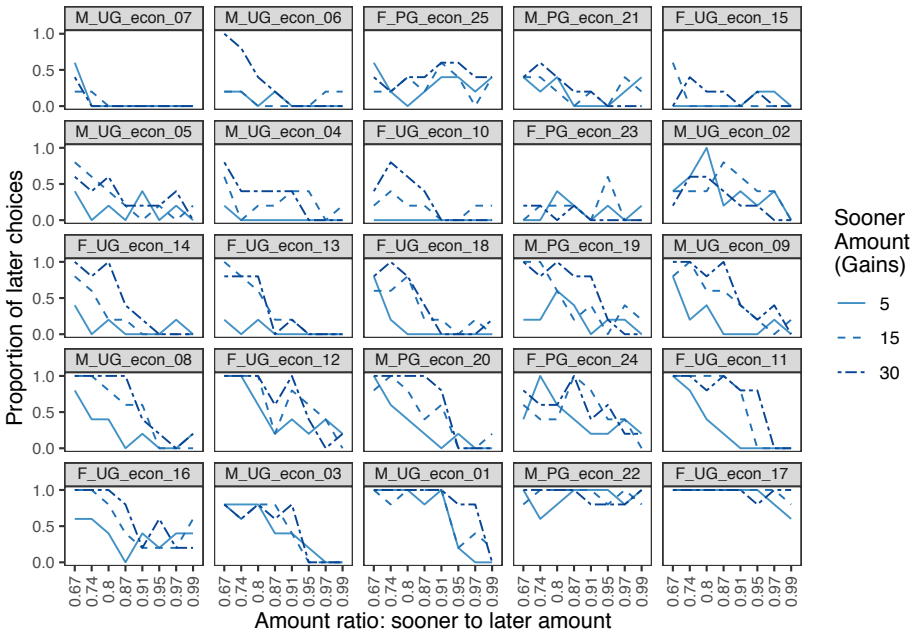


Figure 5.9: Gains only: Line plots showing the proportion of later choices and amount ratio for each participant coloured by the different sooner amounts of money.

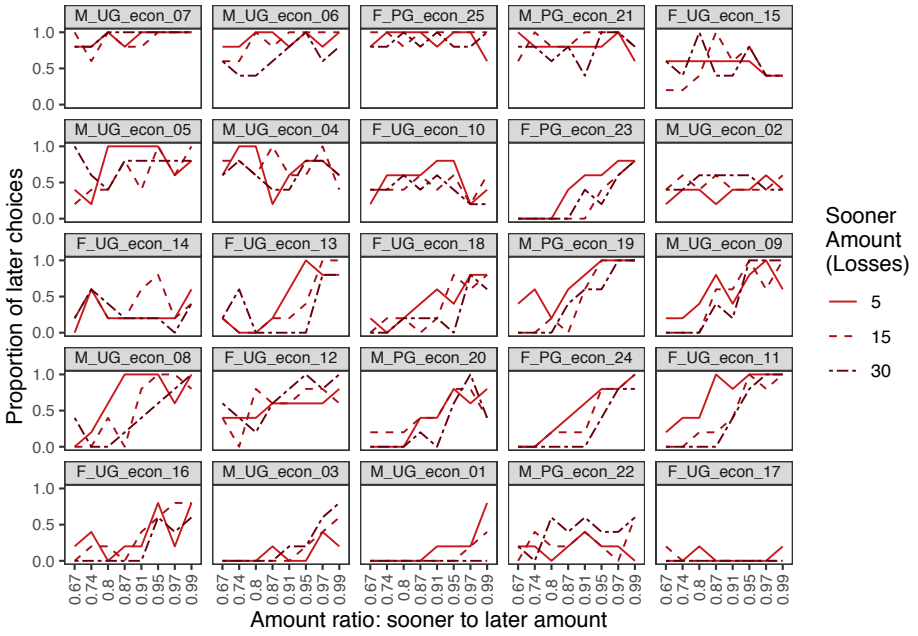


Figure 5.10: Losses only: Line plots showing the proportion of later choices and amount ratio for each participant coloured by the different sooner amounts of money.

5.1.1.4 Indifference points

Figure 5.11 shows the percentage of uncalculable indifference points by sign for each participant. Participants are ordered by increasing percentage of uncalculable indifference points for gains and then losses. Each point, i.e proportion, was based on 30 indifference points, which is based on the factorial design. For example, if the proportion was 0.20 for gains, that means that only 6 indifference points (out of 30) could be calculated. An indifference point is uncalculable if the participant switched more than once or did not switch at all on a unique combination of the factorial design.

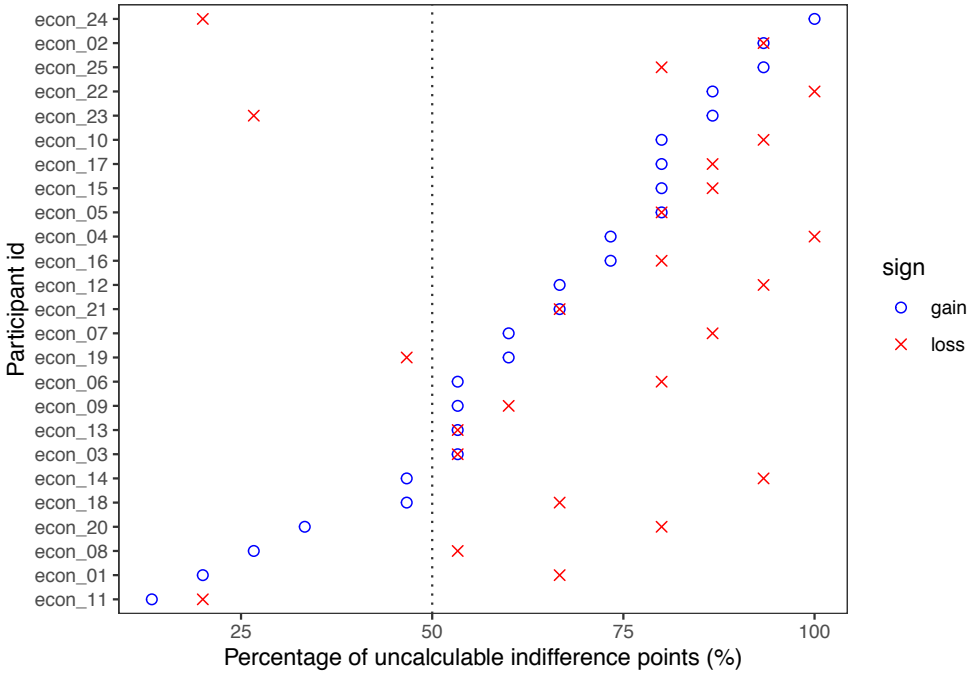


Figure 5.11: Strip chart showing the percentage of indifference points that could not be calculated, coloured by sign for each participant. Participants are ordered by increasing percentage of uncalculable indifference points.

All 25 participants in the study had at least one indifference point (out of 30 possible indifference points) that could not be calculated. The percentange of uncalculable indifference points ranged from 13% to 100% for gains, and 20% to 100% for losses. Only 33% of indifference points could be calculated out of all 750 indifference points

across the 25 participants.

Figure 5.12 shows cumulative density plots of the number of switching points in each unique factorial combination of sign, time delay combination. The panels represent the different time delay combinations and sooner amounts. The lines are coloured by sign.

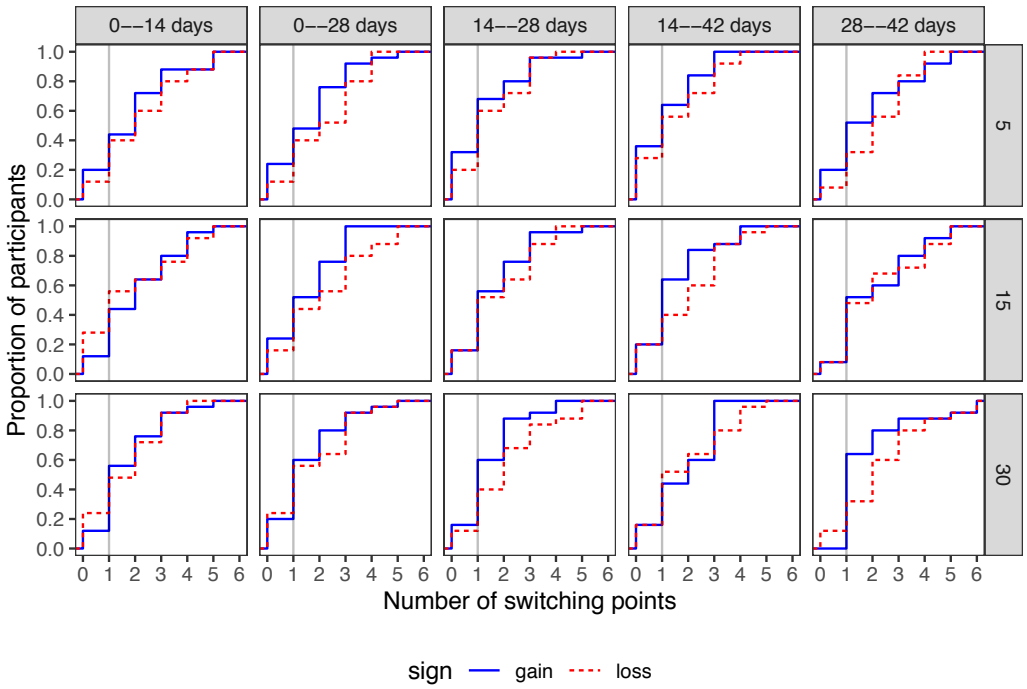


Figure 5.12: Cumulative density plots of the number of switching points for each set of 8 questions from a unique combination of factors. The panels represent the factors, the combination of time delays and the sooner amount. The lines are coloured by sign. A vertical line at 1 on the x-axis is drawn.

There are 30 factors from 2 (sign: gain, loss) × 3 (sooner amounts: 5, 15, 30) × 5 (sooner-later time delay combination: 28-42, 14-42, 14-28, 0-14, 0-28). There were 8 questions on each unique combination of factors. If a participant only switched once on each unique combination of factors, i.e. 8 questions, then the participant would have one unique switching point. If every participant were like this, then the lines on the cumulative density plots would start out flat from zero to one on the x-axis, and then jump up vertically to 1.0 on the y-axis and remain flat across the range of values on the x-axis.

Figure 5.12 shows that for each factorial combination of time delay and sooner amount, participants had different numbers of switching points. The proportion of participants with only a single unique switching point was 0.24 for gains and 0.28 for losses in the first panel on the left. This means that an indifference point could not be estimated for 76% of participants for gains and 72% of participants for losses when the sooner amount was available today and the later amount was available in 2 weeks (14 days). There were 3 participants who switched choices 5 times (on 8 questions) for gains.

Figure 5.13 shows the switching behaviour for each participant when the sooner amount was 5 Euros and the time delay combination was '0–14' days, i.e. the sooner amount was available today and the later amount in 2 weeks. There is substantial heterogeneity in switching behaviours across participants, and even within participants for gains and losses.

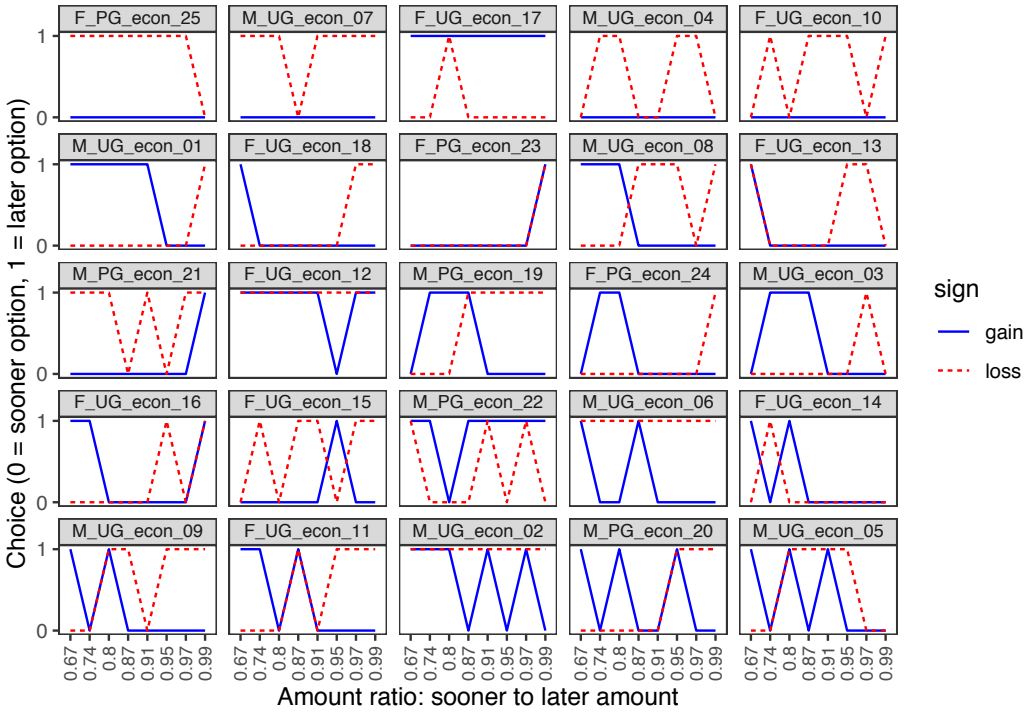


Figure 5.13: Line plots of switching behaviour for each participant when the sooner amount was available today and the later amount in 14 days. Lines are coloured by sign. Participants are ordered by increasing number of switching points for gains and then losses.

5.1.2 Replicating statistical analysis (Faralla et al. 2012)

An attempt was made to replicate the analysis done by Faralla et al. (2012). Faralla et al. (2012) only reported p-values associated with the terms in the logistic regression model without any information on the magnitude or sign of coefficients or the reference categories of the categorical predictors. They had three models: one model contained a subset of data for gains only, another for losses only and the third contained the full data set of gains and losses.

As such, to replicate the analysis by Faralla et al. (2012, 9–10), several assumptions were made due to a lack of information reported. The logistic regression models had choosing the later option as the outcome variable. All three models had the same set of predictors: the sooner amount, interval (later delay minus sooner delay), percentage difference (difference in amounts divided by smaller amount), university student status (undergraduate, postgraduate), age and gender. The models did not include the sign term, i.e. an binary variable to indicate whether the amount offered was a gain or a loss, or any interaction term. The sooner amount was treated as continuous and not categorical. Finally, the reference group for student status is postgraduate and gender is female.

Table 5.3 displays the results from the three models. The first model on the left contains a subset of data for gains only. The second model in the middle contains a subset of data for losses only. The third model on the right contains the full data set.

Overall, the models provide broadly similar results to Faralla et al. (2012) based on p-values with the following exceptions. Some variables were not statistically significant ($p < 0.05$). For example, schooling was not significant in the model with loss-only, which is inconsistent with what was reported by Faralla et al. Finally, it is unclear if the coefficients from our models were in the same direction, or had a similar magnitude, as those in Faralla et al. (2012) who did not report information relating to the coefficients of the terms in their models.

One issue with using a conventional logistic regression model for this data set is

Table 5.3: Table of logistic regression results from attempting to replicate Faralla et al. (2012). The outcome is choosing the later option. The reference categories for schooling and gender are postgraduate students and females. The terms, xs and percentdiff refer to the sooner amount and percentage difference. Models 1 and 2 contain a subset of choices for gains and losses only.

	Regression coefficients: Log odds (95% confidence interval)		
Predictors	Model 1 (gains)	Model 2 (loss)	Model 3 (full)
Intercept	4.56 (3.41, 5.70)	-1.59 (-2.63, -0.55)	1.01 (0.30, 1.72)
xs	0.04 (0.03, 0.04)	-0.01 (-0.02, -0.01)	0.01 (0.00, 0.01)
interval	-0.04 (-0.05, -0.02)	0.03 (0.02, 0.04)	-0.00 (-0.01, 0.01)
percentdiff	5.00 (4.48, 5.52)	-3.73 (-4.23, -3.23)	0.44 (0.12, 0.75)
age	-0.21 (-0.25, -0.17)	0.06 (0.03, 0.10)	-0.05 (-0.08, -0.03)
(schooling)UG	-0.82 (-1.03, -0.61)	-0.03 (-0.22, 0.17)	-0.33 (-0.47, -0.20)
(gender)M	-0.09 (-0.25, 0.08)	0.49 (0.34, 0.64)	0.19 (0.09, 0.29)
Degrees of freedom	2,993	2,993	5,993
AIC	3,418.2	3,843.3	8,150

the repeated measures. Each participant answered 240 questions, which means that responses within a participant are not independent of each other. Treating these repeated responses as independent observations would artificially inflate the sample size and produce inaccurate estimated coefficients. A multilevel model can address this issue.

5.1.2.1 Interpretation of multilevel model

Consider a two-level random intercept logistic model with one predictor variable:

$$\log\left(\frac{\pi_{i,j}}{1-\pi_{i,j}}\right) = \beta_0 + \beta_1 x_{i,j} + u_j$$

where $u_j \sim N(0, \sigma_u^2)$.

As in the single-level model, β_1 is the effect of a 1-unit change in x on the log-odds

when $y = 1$, but in the two-level model it is the effect of x after adjusting for the group effect u. If u is held constant, then x would be the effects for individuals within the same group so β_1 is usually referred to as a cluster-specific effect.

While β_0 is the overall intercept in the linear relationship between the log-odds and x, the intercept for a given group j is $\beta_0 + u_j$, which will be higher or lower than the overall intercept depending on whether u_j is greater or less than zero. As in the continuous response case, u_j is referred to as the group (random) effect, group residual, or level 2 residual. The variance of the intercepts across groups is $\text{var}(u_j) = \sigma_u^2$, which is referred to as the between-group variance adjusted for x, i.e. the level 2 residual variance. See Austin and Merlo (2017) for more details.

The predicted response probability for individual i in group j can be calculated by substituting the estimates of β_0, β_1 and u_j obtained from the fitted model as follows:

$$\hat{\pi}_{i,j} = \frac{\exp(\beta_0 + \beta_1 x_{i,j} + u_j)}{1 + \exp(\beta_0 + \beta_1 x_{i,j} + u_j)}$$

5.1.3 Statistical modelling

The outcome variable would be whether a participant chose the later option on each question, i.e. the outcome is an indicator function that takes the value 1 if the participant chose the later option on that question and 0 if the participant chose the sooner option. The level 1 predictors relate to information at the question-level, e.g. the sign, the sooner and later amounts, the sooner and later delays, and the amount ratio (sooner amount divided by the later amount). The level 2 predictors relate to the information at the participant-level, e.g. gender, education and age.

From the null two-level model estimates (using Laplacian approximation), the log-odds of choosing the later option for an 'average' participant (one with $u_{0j} = 0$) is estimated as $\hat{\beta}_0$ = -0.294. The intercept for participant i is -0.294 + u_{0j}, where the variance of u_{0j} is estimated as $\hat{\sigma}_{u0}^2 = 0.189$.

There is strong evidence that the between-subject variance is non-zero. The likelihood ratio statistic for testing the null hypothesis that the variance of the average

subject is zero, i.e. $\sigma^2_{u0} = 0$, can be calculated by comparing the two-level null model with the correspondong single-level model. The test statistic is 198.7 with 1 degree of freedom from a chi-squared distribution.

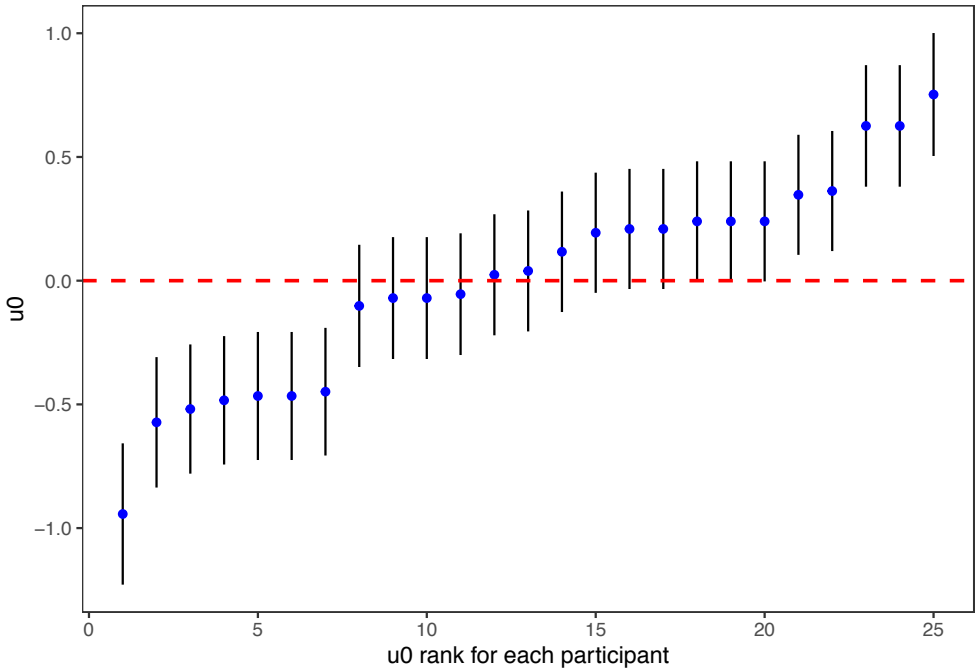


Figure 5.14: Estimated random effects residuals for each participant.

Figure 5.14 shows the estimated residuals for all 25 participants in the sample. For about half of the participants, the 95% confidence interval does not overlap the horizontal line at zero, indicating that the choices for the later option are significantly above or below average (above/below the zero line). The confidence intervals have the same width because each participant answered the same number of questions (240 questions).

Results from 7 models are displayed in Table 5.4. The terms in the models were selected based on the findings from the exploratory data analysis. The estimated coefficients are rounded to two decimal places. The reference groups for the categorical variables are: sign: gain; interval: 14 days; the sooner amount: 5 Euros. Level 2 predictors did not have lower AIC values compared to the null model. Models with age and an interaction between schooling and gender failed to converge.

Table 5.4: Table of multilevel logistic regression results.

	Regression coefficients: Log odds (95% confidence interval)						
Terms	Null model	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Intercept	-0.29 (-0.47, -0.12)	-0.49 (-0.68, -0.30)	0.07 (-0.38, 0.53)	-0.06 (-0.52, 0.40)	0.08 (-0.38, 0.54)	-0.39 (-0.86, 0.08)	-0.38 (-0.86, 0.09)
(sign)loss		0.39 (0.28, 0.49)	0.39 (0.28, 0.49)	0.39 (0.29, 0.50)	0.39 (0.28, 0.49)	1.03 (0.84, 1.22)	1.03 (0.84, 1.22)
amount ratio			-0.65 (-1.13, -0.18)	-0.66 (-1.14, -0.18)	-0.65 (-1.13, -0.18)	-0.66 (-1.15, -0.18)	-0.66 (-1.15, -0.18)
(xs)15				0.18 (0.05, 0.31)		0.57 (0.38, 0.76)	0.57 (0.38, 0.76)
(xs)30				0.21 (0.08, 0.34)		0.81 (0.62, 1.00)	0.81 (0.62, 1.00)
(interval)28					-0.02 (-0.13, 0.09)		-0.02 (-0.13, 0.09)
(sign)loss:(xs)15						-0.73 (-1.00, -0.47)	-0.73 (-1.00, -0.47)
(sign)loss:(xs)30						-1.14 (-1.40, -0.88)	-1.14 (-1.40, -0.88)
Random effects							
τ_{00}	0.19	0.19	0.19	0.19	0.19	0.20	0.20
ICC	0.05	0.06	0.06	0.06	0.06	0.06	0.06
Participants	25	25	25	25	25	25	25
Observations	6,000	6,000	6,000	6,000	6,000	6,000	6,000
AIC	8,006.8	7,955.7	7,950.6	7,942.7	7,952.5	7,870.8	7,872.6

Reference categories for sign is gain, xs is 5 Euros, and interval is 14 days. The term, xs, refers to the sooner amount.

Across the models, the intraclass correlation coefficient (ICC), also known as the the variance partition coefficient, is about 0.06. This means about 6% of the variance is explained by the grouping structure in the population. In Table 5.4, τ_{00} represents the between-subject variance.

Model 5 had the lowest AIC of 7870.8 with the terms: sign, amount ratio, the sooner amount and the interaction between sign and the sooner amount. Adding the interaction term in the model improved the AIC substantially compared to the model without the interaction. However, adding in the interval term in a subsequent model did not improve the AIC. Similarly, although not shown, adding in level 2 predictors (age or schooling) did not improve the AIC.

5.1.4 Predicted probabilities

The plots below show the relationship between the predicted probabilities of choosing the later option and the later amount. The relationship is displayed separately in different panels for sign and time delay interval. The points and lines are coloured by the different sooner amounts. Only results from selected models are shown.

In the initial models, the predicted probabilities for choosing the later gain are mostly below 0.5. The probabilities for the later losses tend to be slightly higher than the later gains but this difference becomes smaller as more terms are included. Within each panel, there are 4 distinct groups of participants, e.g. there are a few participants at the top, a bigger group of participants bunched in the middle, a smaller group below and one participant right at the bottom.

The plots for the null model and model 1 are similar in that the predicted probabilities are constant for each individual. However, in model 1, the predicted probabilities for losses tend to be shifted higher than for gains. The plots for models 2, 3 and 4 are similar with some differences in the steepness of the lines between the sooner amounts (e.g. when the red lines join the green lines, which then join the blue lines). Finally, the plots for models 5 and 6 are similar. They differ from the plots for models 2, 3 and 4 in that the gradient of the lines for gains tend to be positive between

the sooner amount values compared to losses.

The predicted probabilities for Model 5 are similar to Model 6 as the only difference between the two models is that Model 6 contains an additional interval term, which has an odds ratio of 1.0 and did not add significantly to Model 5. There are four distinct clusters of participants' responses in Model 5. The model predicts that participants are more likely to choose the sooner gain, as most points for gains lie below the 0.5 line, and the later loss.

For gains, the predicted probabilities of choosing the later option increases monotonically as the later amount increases. For losses, the predicted probabilities increase as the later amount increases within each sooner amount but decreases as the sooner amount increases. These patterns are consistent with the relationship where the proportion of later choices decreases for gains as the amount ratio increases for gains but increases for losses.

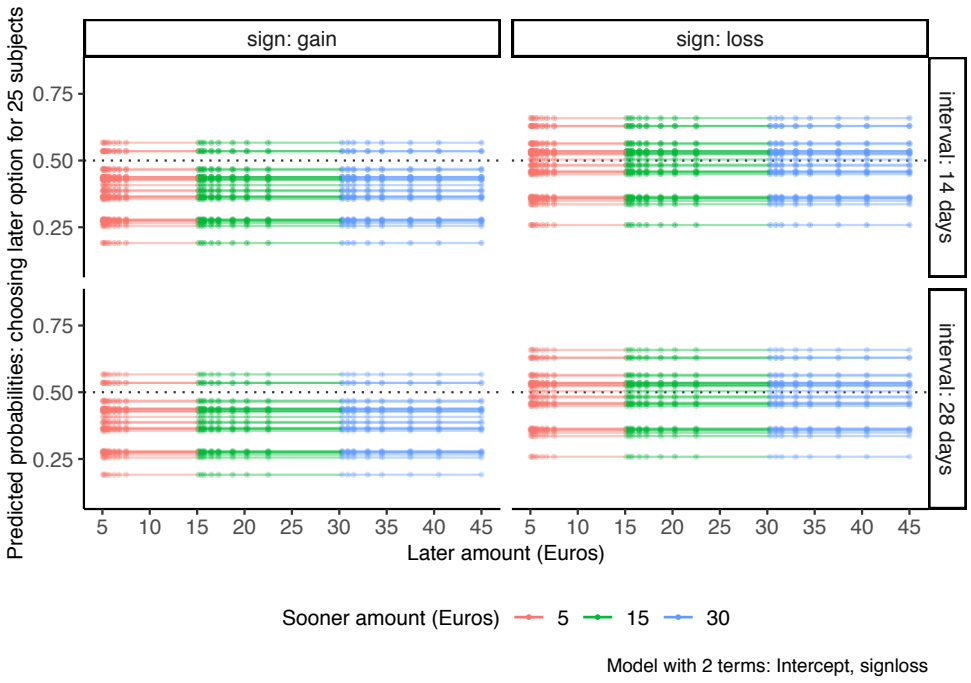


Figure 5.15: Predicted probabilities for model 1.

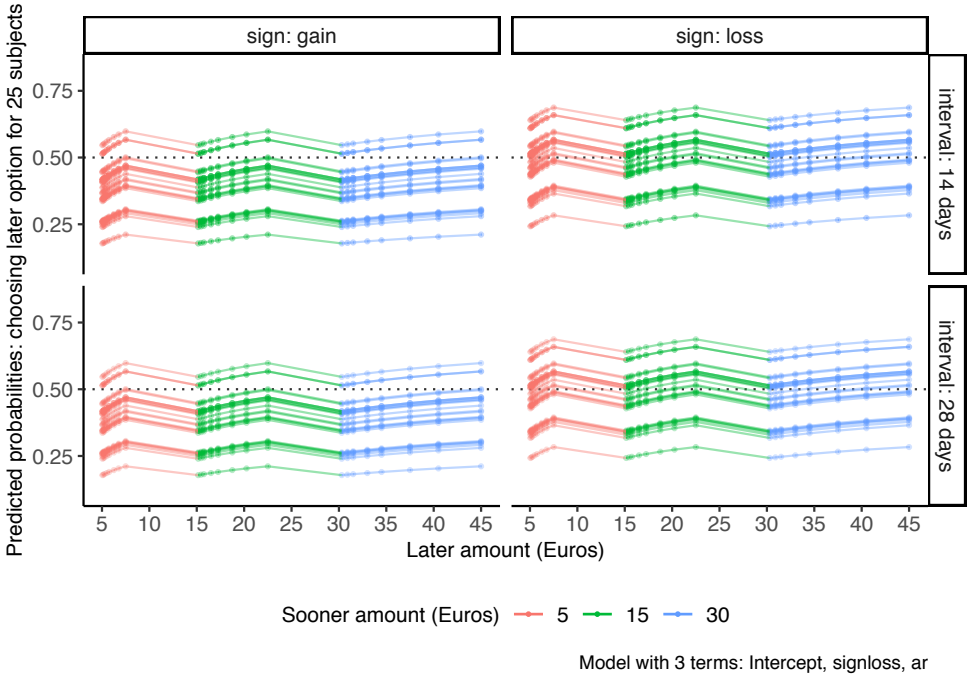


Figure 5.16: Predicted probabilities for model 2.

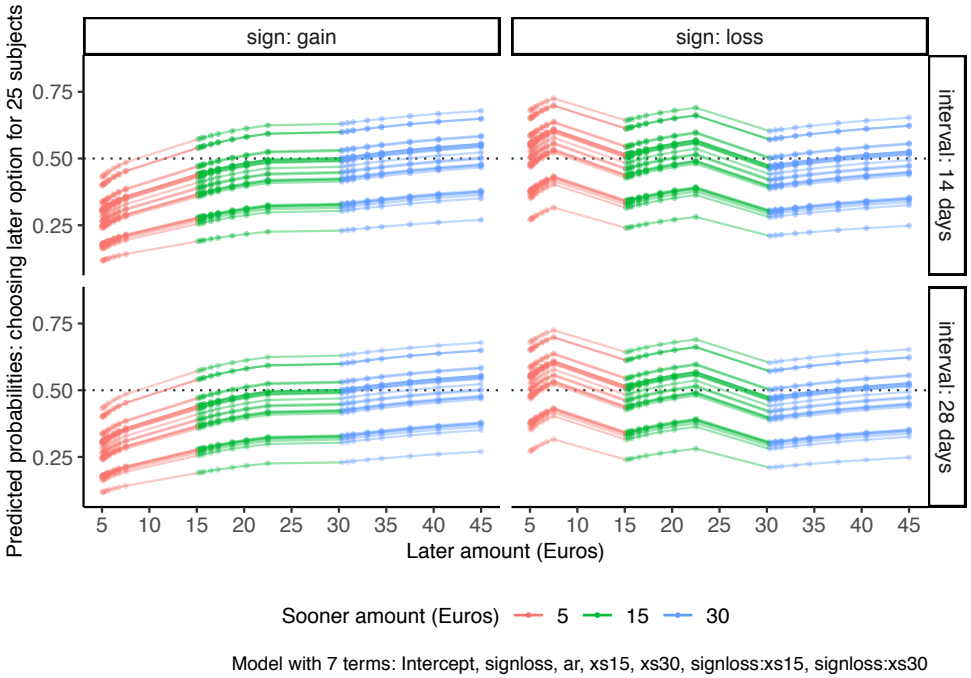


Figure 5.17: Predicted probabilities for model 5.

5.1.5 Diagnostics

Additional diagnostics can be assessed (D. G. Altman and Bland 1994a, 1994b), e.g. sensitivity, specificity, and predictive values. Sensitivity is the 'proportion of true positives that are correctly identified'. Specificity is the 'proportion of true negatives that are correctly identified'. The positive predictive value (PPV) refers to the proportion of correctly identified later choices. The negative predictive value (NPV) refers to the proportion of correctly identified sooner choices.

The table below presents the diagnostic results for model 2. It provides an additional assessment of the model. The results are separated for gain questions only, loss questions only, and for all questions (both gain and loss).

Table 5.5: Diagnostics results for model 5. The later choice is taken as a "positive", while the sooner choice, a "negative".

Sign	Later choices	Sooner choices	Sensitivity (%)	Specificity (%)	Prevalence (%)	PPV (%)	NPV (%)
all	2,583	3,417	43.4	76.1	43.0	57.8	64.0
gain	1,155	1,845	29.2	87.6	38.5	59.6	66.4
loss	1,428	1,572	55.0	62.5	47.6	57.1	60.4

Overall, model 5 is able to correctly predict the later choice less than half (44%) of the time (sensitivity) and the sooner choice about three-quarters of the time (specificity). The model is able to correctly predict the later gain less than a third (29%) of the time (sensitivity for gain) but the sooner gain almost 90% of the time (specificity for gain). For losses, the model correctly predicts the later loss 55% of the time and the sooner loss 62% of the time.

For the given prevalence, overall, the model is able to correctly predict the sooner (NPV) and later (PPV) choices about 64% and 58% of the time respectively. The model is able to correctly predict the sooner and later gains about 66% and 60% of the time respectively. The model is able to correctly predict the sooner and later losses about 60% of the time.

5.2 Xu et al. 2009

This section focusses on Xu et al. 2009. Xu et al. 2009 did not explicitly mention the study design employed but reported that 'the percent difference in amounts between the two rewards was selected from the set {5%, 10%, 15%, 25%, 35%, 50%}'. One could reasonably assume an implicit within-participant factorial design: 2 (sign: gain, loss) x 5 (percent difference: 5, 10, 15, 25, 35, 50). However, since the 'percent difference' was not explictly defined, calculating it as the difference in the sooner and later amount divided by either the sooner or later amount did not yield a result with 5 unique values, e.g. when the sooner amount is the denominator there are 14 unique values after rounding to the nearest whole number: 5, 10, 14, 15, 17, 25, 26, 34, 35, 36, 50, 51, 52, 54.

The study design could also incorporate the time delays. For example, there are 6 unique combinations: 0-14 days, 0-28 days, 14-28 days, 14-42 days, 28-42 days, 28-56 days, where the dash separates the time to the sooner delay and the time to the later delay. The interval, i.e. difference in time delays, has 2 unique values: 14 days, 28 days.

There were 22 participants in Xu et al. 2009's fMRI study but two 'were excluded from the analysis because of excessive head motion'. In this data set, there are 20 participants. Data for the two excluded participants were not provided to us.

5.2.1 Exploratory data analysis

Figure 5.18 is an attempt to check for any potential coding errors introduced from cleaning and manipulating the raw data for this analysis. It closely reproduces Figure 1 from Xu et al. 2009, which showed the proportion of choices by sign. The proportion of early gains and losses are 0.5 and 0.76 respectively.

There were 10 women and 10 men. They were 'right-handed Chinese graduate students' with a mean age of 25 years (range: 22 to 29 years). Gender was the only demographics information provided in the dataset.

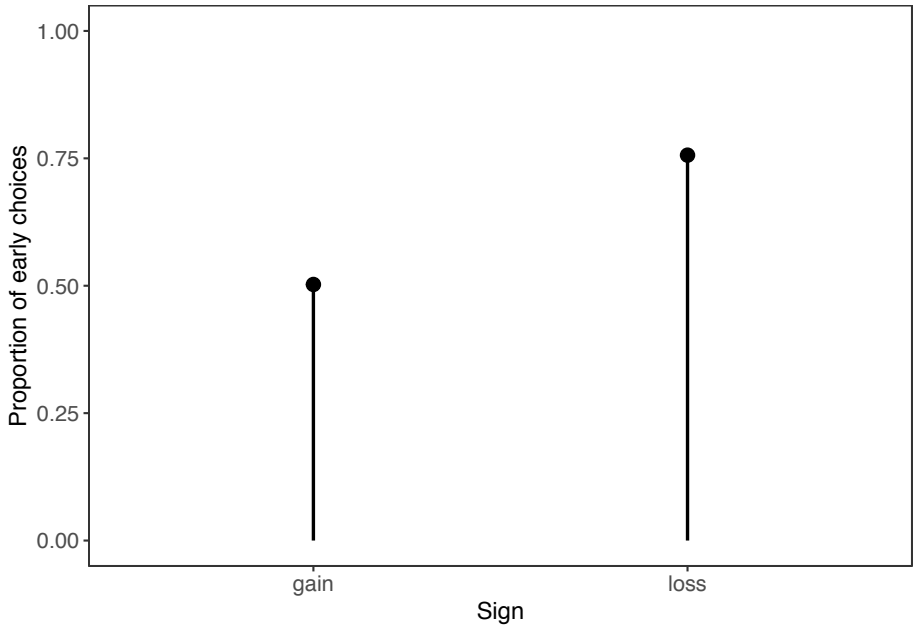


Figure 5.18: The proportion of sooner choices for gains and losses across all participants, reproducing Figure 1 from Xu et al. 2009.

Each participant was asked 40 questions, half of which were gains and half were losses. The sooner amounts offered ranged from CNY 13 to CNY 110. The later amounts offered ranged from CNY 20 to CNY 149.

Table 5.6 shows for the different combination of sooner and later delays the number of unique: questions asked per participant; differences in amounts in the questions (later amount minus sooner amount); and amount ratios in the questions (sooner amount divided by the later amount). For example, there were 8 questions with the sooner amount available now and the later amount in 2 weeks (14 days). These 8 questions had 8 unique amount difference and 6 unique amount ratios.

The number of unique amount difference and unique amount ratios are not always equal to the number of questions. This reflects the study design. One implication of such a study design is that an indifference point cannot be estimated accurately for every unique combination of the sooner and later time delays, i.e. for those where the number of questions and number of amount ratio or difference are not the same.

Table 5.6: Breakdown of the number of questions by the different time delays for each domain of sign (gain/loss). The table is ordered by the difference in delays and then by the combination.

Combination of sooner and later delays (days)	Difference in delays (days)	Questions (gain/loss)	Unique amount ratios	Unique amount difference
0-14	14	8	6	8
14-28	14	5	5	5
28-42	14	6	6	4
0-28	28	8	7	6
14-42	28	7	7	7
28-56	28	6	6	6

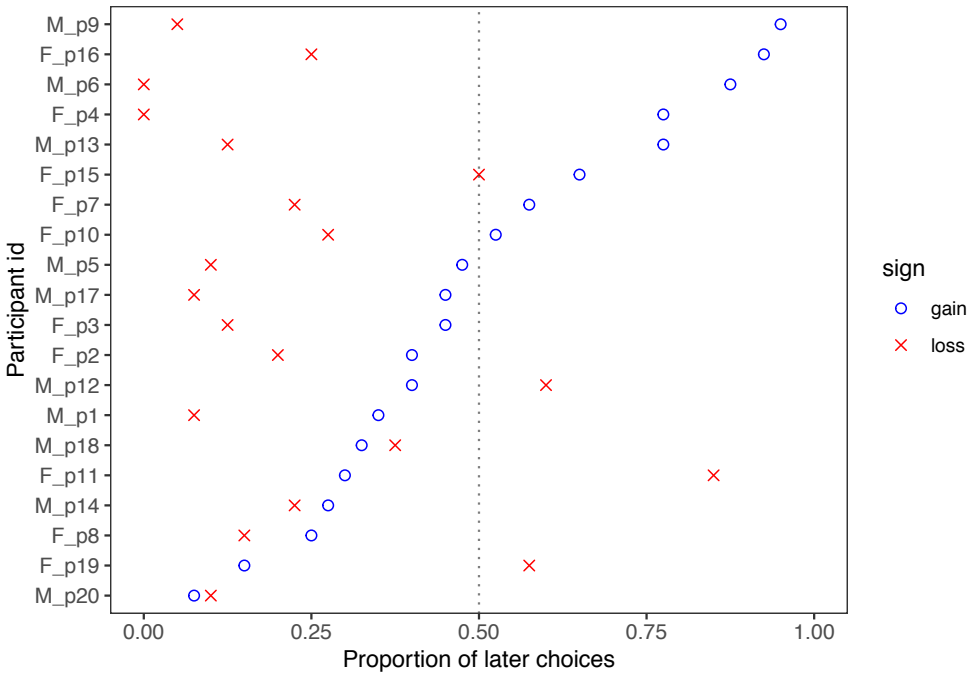


Figure 5.19: Strip chart showing the proportion of later choices for each participant. The participants on the x-axis are ordered by increasing proportion of later gains. The colour and shape of the points represent the two different values of sign: gain and loss.

Figure 5.19 shows the proportion of later choices for each participant. Each point represents the overall proportion of later choices for one participant. The colour and shape of the points indicate whether the proportion is for gains (blue circle) or losses

(red crosses). There is a dotted vertical line at 0.5 on the x-axis to highlight the points to the left and right of the line.

From Figure 5.19, most (70%) proportions are smaller than 0.5, which implies that most choices are for the sooner option. Only 3 points for losses are greater than 0.5 compared to 8 points for gains. The median proportion of later gains is higher than that of losses (0.45 vs. 0.18). Most (75%) points for gains are to the right of losses within participants. Women tend to choose the later loss more often than men (median: 0.24 vs. 0.1). For gains, women tend to choose the later option slightly more than men (median: 0.49 vs. 0.43).

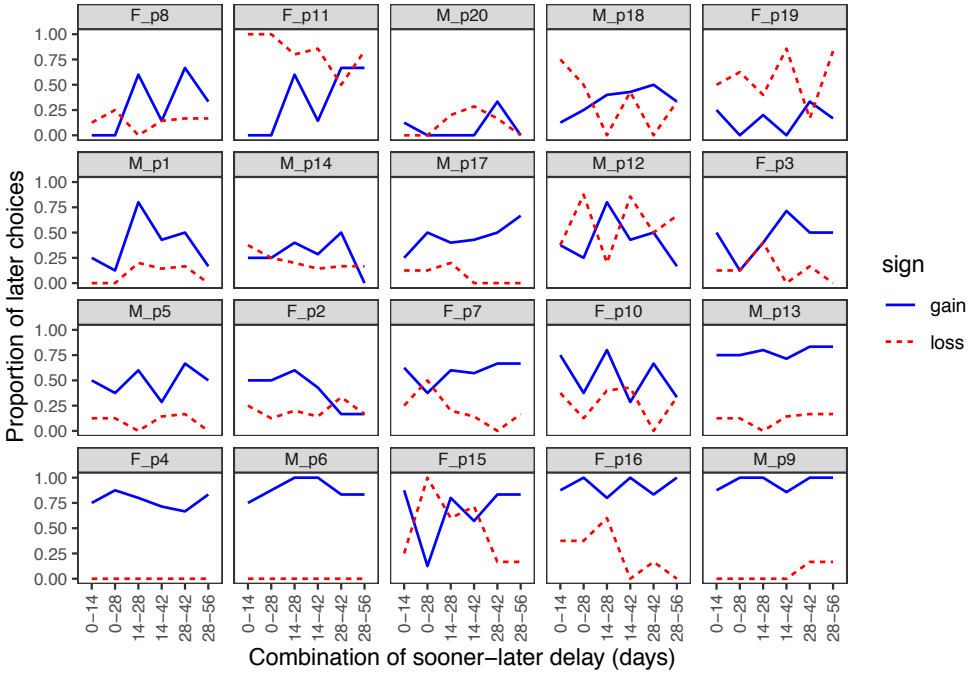


Figure 5.20: Line plots of the relationship between the proportion of later choices and the time delays for each participant. Lines are coloured by sign. Each panel heading provides information on the gender and study id for each participant.

5.2.1.0.1 Time delay Figure 5.20 shows the relationship between the proportion of later choices and time delay for the sooner and later option, for each participant. Participants are ordered by increasing proportion of later gains at each time delay.

Figure 5.21 shows line plots of the relationship between the proportion of later choices

and time interval split by sign and the pattern of relationship. For example, the top left panel shows the lines for participants who had a smaller proportion of later gains when the interval was 28 days compared to 14 days.

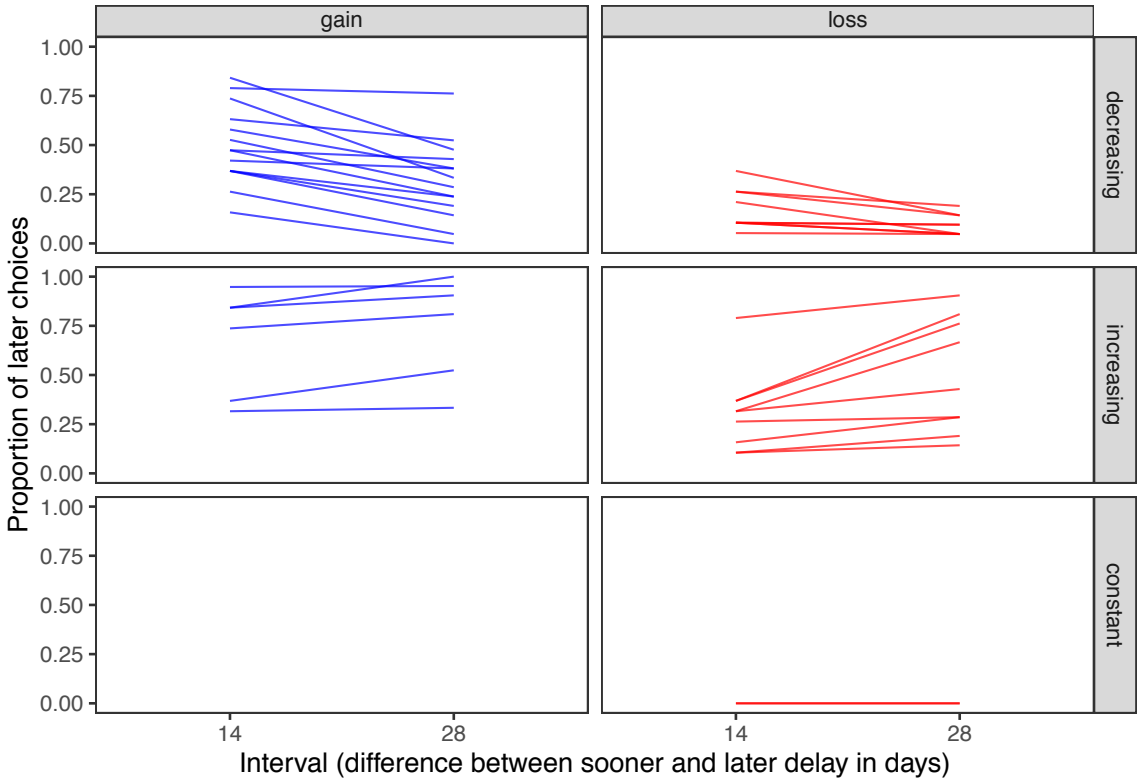


Figure 5.21: Line plots of the relationship between the proportion choosing the later option and the delay interval split by sign and whether the pattern was strictly decreasing, strictly increasing or constant over the interval values.

For gains, 70% of the 20 participants (6 men and 8 women) had a greater proportion of later choices when the interval was 2 weeks (14 days) compared to 1 month (28 days). This is represented in the top left panel in the Figure above. For losses, 45% of the 20 participants (6 men and 3 women) had a greater proportion of later choices when the interval was 2 weeks (14 days) compared to 1 month (28 days). This is represented in the top right panel in the Figure above.

Figure 5.22 shows the relationship between the proportion of later choices (y-axis) and gender (x-axis) by interval and gain. The median proportion of later choices is higher for gains than losses. For gains, the proportion of later choices is slightly

higher when the interval is 14 days compared to 28 days but there is no discernable gender differences within each interval. For losses, there is no discernable differences by interval but within an interval, the proportion of later choices for women is slightly higer than for men.

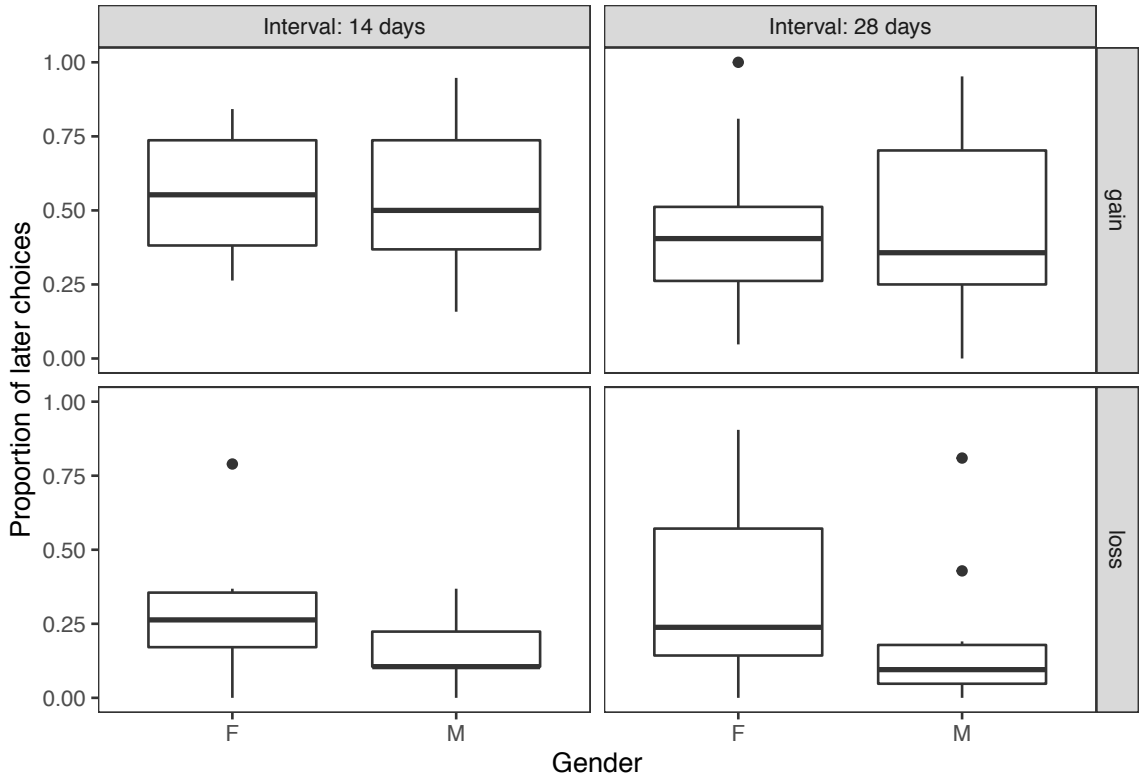


Figure 5.22: Boxplots of the proportion of later choices and gender by interval and sign.

5.2.1.0.2 Money amounts Table 5.7 shows the number of questions for each unique amount ratio rounded to two decimal places. The amount ratio is defined as the sooner amount divided by the later amount. It is rounded to get the proportion of later choices for the purposes of exploring the data.

There were 12 unique values of the amount ratio rounded to two decimal places. The number of questions with each rounded amount ratio ranged from 1 to 8. There were 5 rounded amount ratios with only one question. For these rounded amount ratios, the proportion of sooner or later choice for each participant would be either 0 or 1.

Table 5.7: Number of questions for each unique value of amount ratio rounded to two decimal places.

Amount ratio (rounded to 2 d.p.)	Number of questions
0.65	1
0.66	5
0.67	7
0.73	1
0.74	8
0.79	4
0.80	5
0.86	1
0.87	3
0.88	1
0.91	3
0.95	1

Figure 5.23 shows the relationship between the proportion of later choices and amount ratio by sign. The lines have a different linetype depending on the time interval. As the amount ratio increases, the proportion of later gains decreases while the proportion of later losses increases. This relationship is similar across the different time intervals.

Figure 5.24 shows the relationship between the proportion of later choices and amount ratio, rounded to two decimal places, for each participant. The size of the points represents the number of questions that have the unique value of amount ratio (rounded to 2 decimal places). The number of questions range from 1 to 8.

Compared to the same plots in other studies, the lines in this study fluctuate more. For the rounded amount ratios with only one question, the proportion of sooner or later choice for each participant would be either 0 or 1.

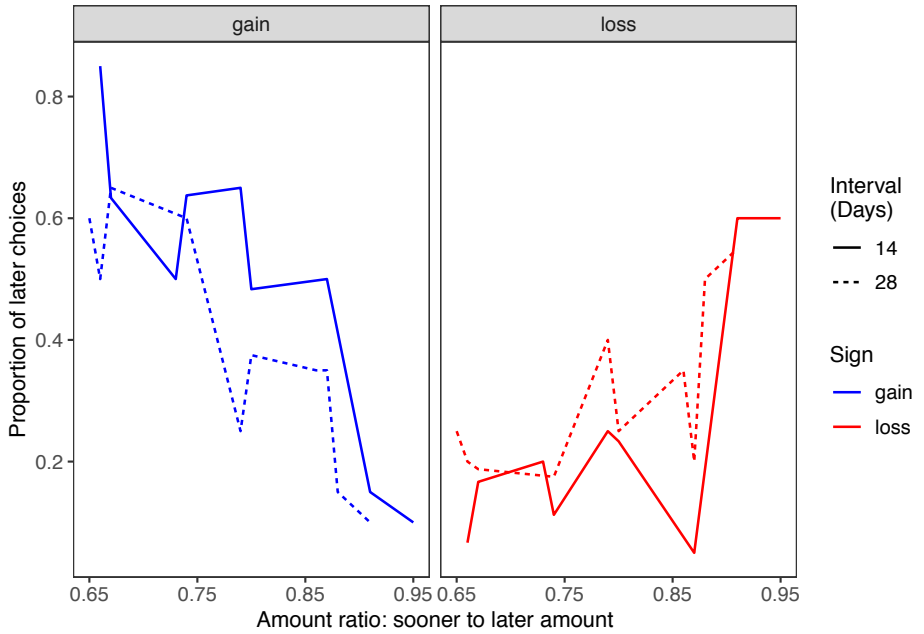


Figure 5.23: Line plots of the proportion of later choices and amount ratio by sign. Lines have different patterns depending on the time interval.

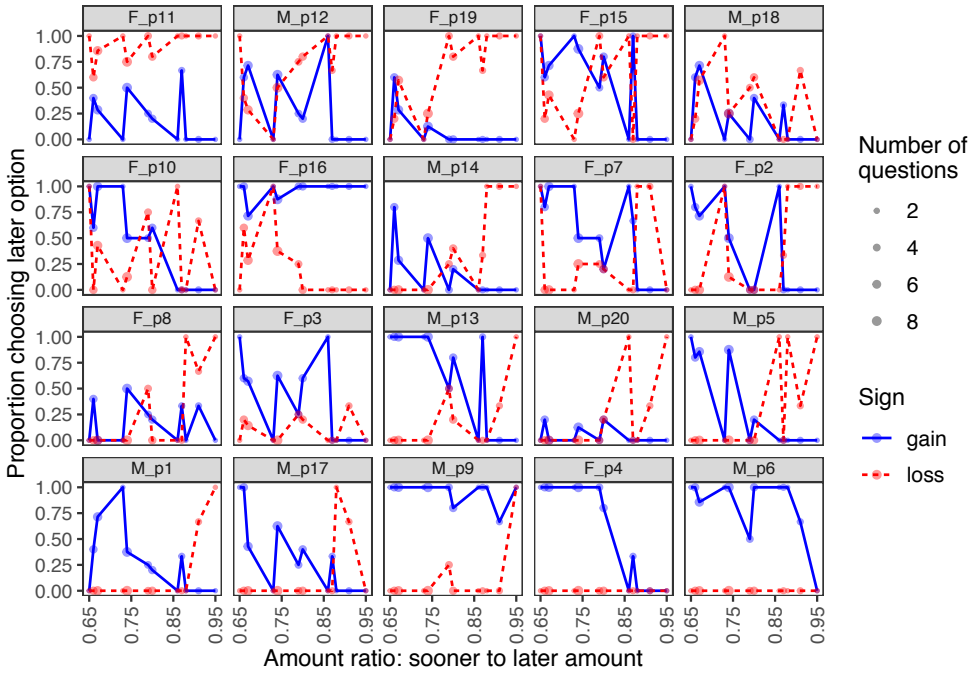


Figure 5.24: Line plots of the relationship between the proportion of later choices and amount ratio (rounded to 2 d.p.) for each participant.

5.2.1.1 Indifference points

A unique indifference point could not be accurately estimated for each participant on some combinations of the factorial design because there were multiple questions with the same unique amount ratio on some factorial combinations. If a participant had different choices for questions with the same amount ratio on a given factorial combination, e.g. by choosing the sooner option on one question and the later option on another question, then it is unclear if either, or which, response should be taken as a switching point.

Instead of using the amount ratio, the amount difference would be used to estimate indifference points for participants. The amount difference, defined as the sooner amount subtracted from the later amount, had fewer repeated unique values. Table 5.8 shows the number of questions and duplicate unique values of amount difference for each of the 12 possible combinations from the factorial design of of sign, and sooner and delay combination. For each of the two sooner-later delay combinations, there was one duplicate unique values of amount difference. Indifference points will only be estimated for the four combinations where there are no duplicate unique amount differences.

Table 5.8: Number of questions and duplicated values of amount difference for each unique factorial combination.

Sooner-later delay combination (days)	No. duplicate unique amount difference values	No. gain questions	No. loss questions
0-14	0	8	8
0-28	1	6	6
14-28	0	5	5
14-42	0	7	7
28-42	1	4	4
28-56	0	6	6

Figure 5.25 shows the percentage of uncalculable indifference points by sign for each participant. Participants are ordered by increasing percentage of uncalculable indif-

ference points for gains and then losses. Each point, i.e proportion, was based on 4 indifference points, i.e. one indifference point for each time delay combination that did not have repeated unique amount differences. For example, if the proportion was 0.25 for gains, that means that only 1 indifference point (out of 4) could be calculated.

All 20 participants in the study had at least one indifference point (out of 4 possible indifference points) that could not be calculated. The percentange of uncalculable indifference points ranged from 0% to 100% for gains and 25% to 100% for losses. Only 31% of indifference points could be calculated out of all 160 indifference points across the 20 participants.

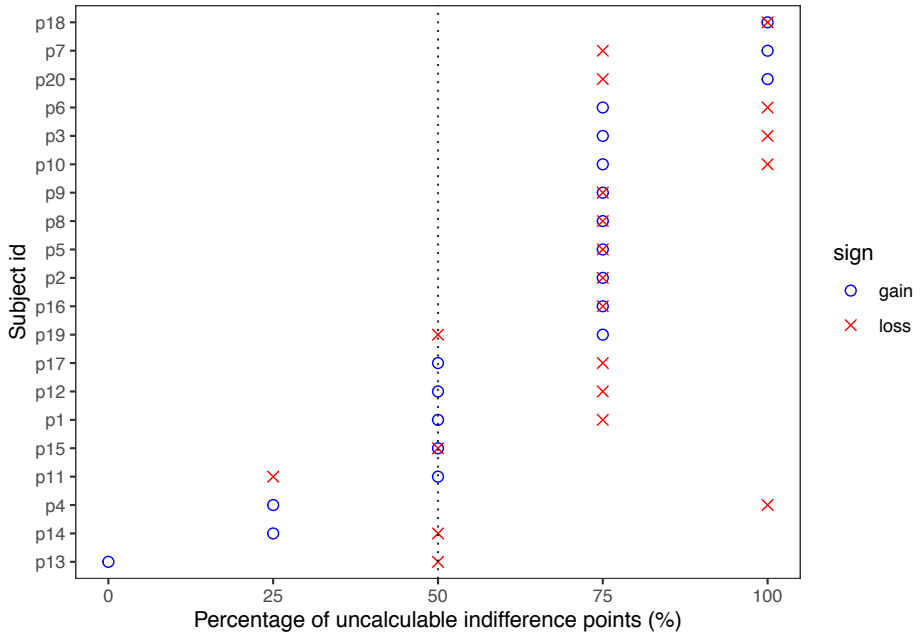


Figure 5.25: Percentage of indifference points that could not be calculated coloured by sign for each participant. Participants are ordered by increasing percentage of uncalculable indifference points.

Figure 5.26 shows cumulative density plots of the number of switching points in each unique factorial combination of sign, time delay combination. The panels represent the different time delay combinations and the lines are coloured by sign.

If a participant only switched once on each of the 8 unique factorial combinations,

then a participant would have 8 unique switching points in total, 4 for gains and 4 for losses. If every participant were like this, then the lines on the cumulative density plots would start out flat from zero to one on the x-axis, and then jump up vertically to 1.0 on the y-axis and remain flat across the range of values on the x-axis.

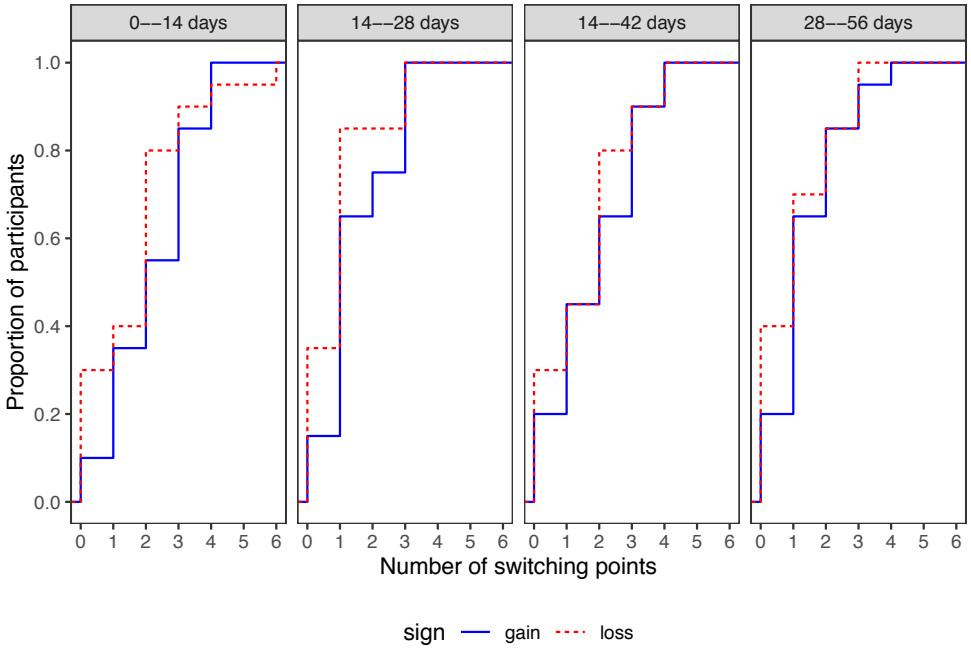


Figure 5.26: Cumulative density plots of the number of switching points in each unique factorial combination of sign and time delay.

Figure 5.26 shows that at each time delay combination, participants had different numbers of switching points. The proportion of participants with only a single unique switching point was 0.25 for gains and 0.1 for losses in the first panel on the left. This means that an indifference point could not be estimated for 75% of participants for gains and 90% of participants for losses when the sooner amount was available today and the later amount was available in 2 weeks (14 days). There was one participant who switched six times on 8 questions for losses.

Figure 5.27 shows the switching behaviour for each participant when the time delay combination was '0–14' days, i.e. the sooner amount was available today and the later amount in 2 weeks. There is substantial heterogeneity in switching behaviours across participants. The participant in the bottom right panel (F_p10) is the participant

with 6 switching points for losses in the previous Figure.

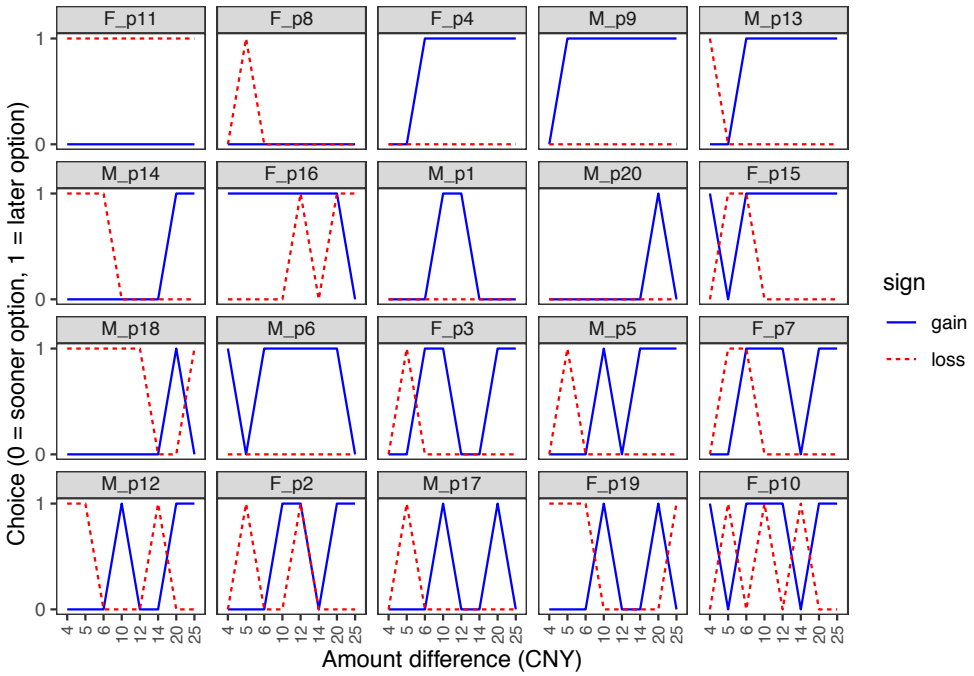


Figure 5.27: Line plots of switching behaviour for each participant when the sooner amount was available today and the later amount in 14 days. Lines are coloured by sign. Participants are ordered by increasing number of switching points for gains and then losses.

5.2.2 Statistical modelling

There is evidence that the between-subject variance is non-zero. The likelihood ratio statistic for testing the null hypothesis that the variance of the average subject is zero, i.e. $\sigma^2_{u0} = 0$, can be calculated by comparing the two-level null model with the correspondong single-level model. The test statistic is 70.2 with 1 degree of freedom from a chi-squared distribution.

Figure 5.28 shows the estimated residuals for all 20 participants in the sample. Eight participants have a 95% confidence interval that does not overlap the horizontal line at zero. For these participants, the choices for the later option are significantly above or below average (above/below the zero line).

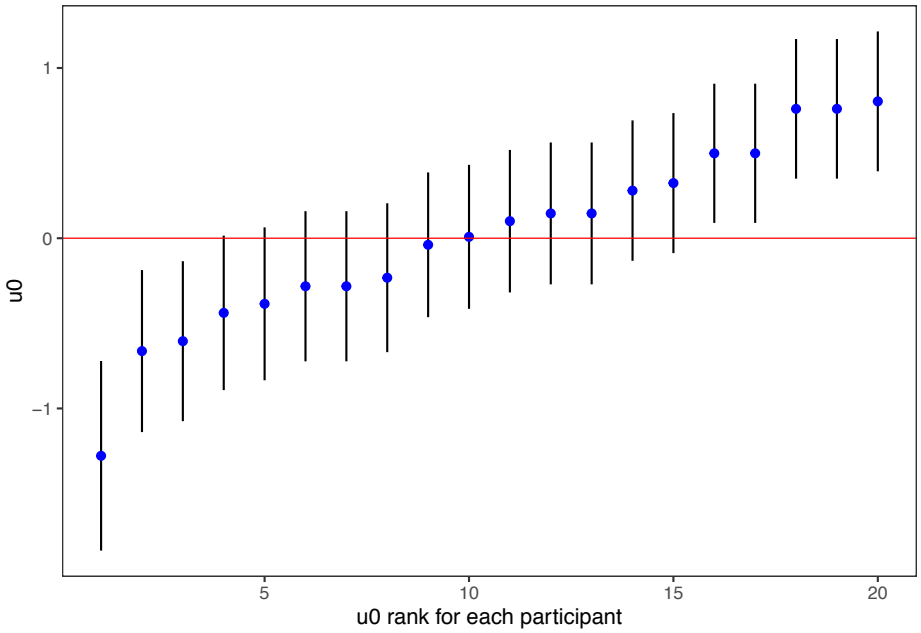


Figure 5.28: Estimated random effects residuals for each participant.

Results from 6 models are displayed in Table 5.9. The models were selected based on the findings from the exploratory data analysis. The estimated coefficients are rounded to two decimal places for presentation purposes. Sign, gender and interval are categorical variables. The reference category for sign is gain, for gender is women and for interval is 14 days.

Across the models, the intraclass correlation coefficient (ICC) ranges between 0.09 and 0.11. This means about 9–11% of the variance is explained by the grouping structure in the population. In Table 5.9, τ_{00} represents the between-subject variance.

The first four models, after the null model, included terms based on the EDA findings. The last model included interval, which did not lower the AIC value. This aligns with the EDA, which did not show any discernable relationship with the proportion of later choices. There were substantial reductions in the AIC values when sign was added. For example, model 1 had 2 terms (intercept, sign) and an AIC of 1,925 compared to the null model, which had an AIC of 2,044. Models with an interaction between amount ratio and sign produced unreliable results, i.e. very large coefficients.

Table 5.9: Table of multilevel logistic regression results.

	Regression coefficients: Log odds (95% confidence interval)					
Terms	Null model	Model 1	Model 2	Model 3	Model 4	Model 5
Intercept	-0.57 (-0.85, -0.30)	-0.02 (-0.33, 0.29)	1.40 (0.38, 2.41)	1.59 (0.53, 2.64)	1.42 (0.36, 2.49)	1.66 (0.54, 2.78)
(sign)loss		-1.21 (-1.43, -0.99)	-1.22 (-1.44, -0.99)	-1.22 (-1.44, -0.99)	-0.83 (-1.13, -0.54)	-0.84 (-1.14, -0.54)
amount ratio			-1.86 (-3.13, -0.59)	-1.86 (-3.13, -0.59)	-1.87 (-3.14, -0.59)	-1.89 (-3.16, -0.61)
(gender)M				-0.38 (-0.96, 0.19)	-0.03 (-0.64, 0.58)	-0.03 (-0.64, 0.58)
(sign)loss:(gender)M					-0.85 (-1.30, -0.39)	-0.85 (-1.30, -0.39)
interval						-0.01 (-0.03, 0.01)
Random effects						
τ_{00}	0.33	0.39	0.40	0.36	0.37	0.38
ICC	0.09	0.11	0.11	0.10	0.10	0.10
Participants	20	20	20	20	20	20
Observations	1,600	1,600	1,600	1,600	1,600	1,600
AIC	2,043.6	1,925.0	1,918.8	1,919.1	1,907.6	1,907.8

Reference categories for gender is female and for sign is gain.

5.2.3 Predicted probabilities

The plots below show the relationship between the predicted probabilities of choosing the later option and the amount ratio (defined as the sooner amount divided by the larger amount). The points and lines are coloured by gender.

Across all models, the predicted probabilities for choosing the later loss are almost always below 0.5. As more terms are included in the model, the predicted probabilities for choosing the later gain and loss tend to increase for each participant

The relationship between the predicted probabilities and amount ratio is constant for the null model. The model with only the sign term shifts the predicted probabilities up for gains and down for losses. Once amount ratio is included (third model), then the relationship is negative, with some slopes steeper than others. Once the interaction between sign and gender is included, gender differences become more apparent, e.g. women tend to have higher predicted probabilities for choosing the later losses.

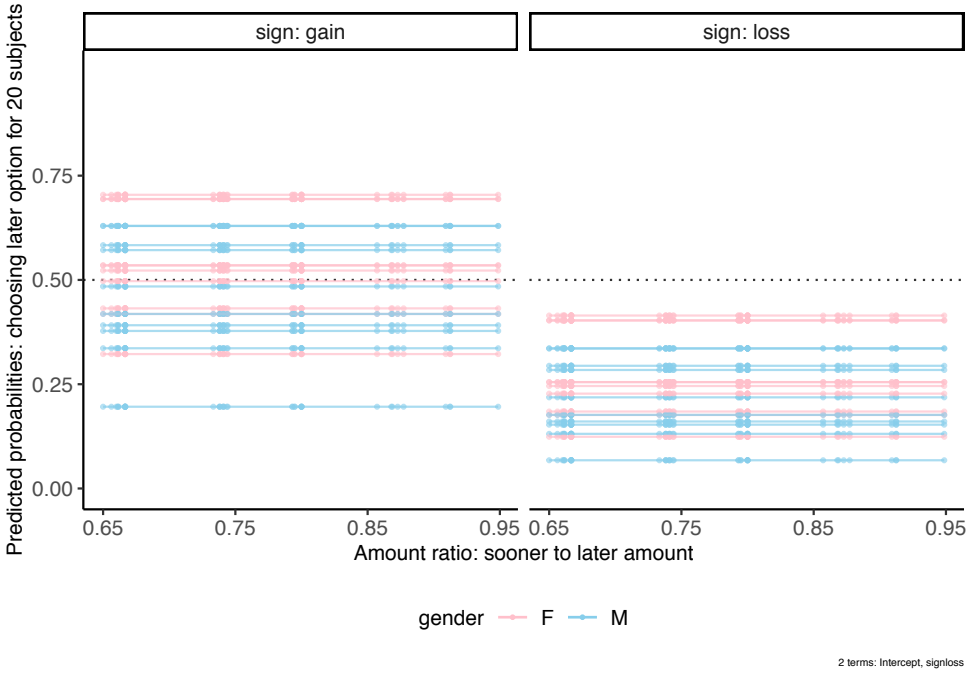


Figure 5.29: Predicted probabilities for model 1.

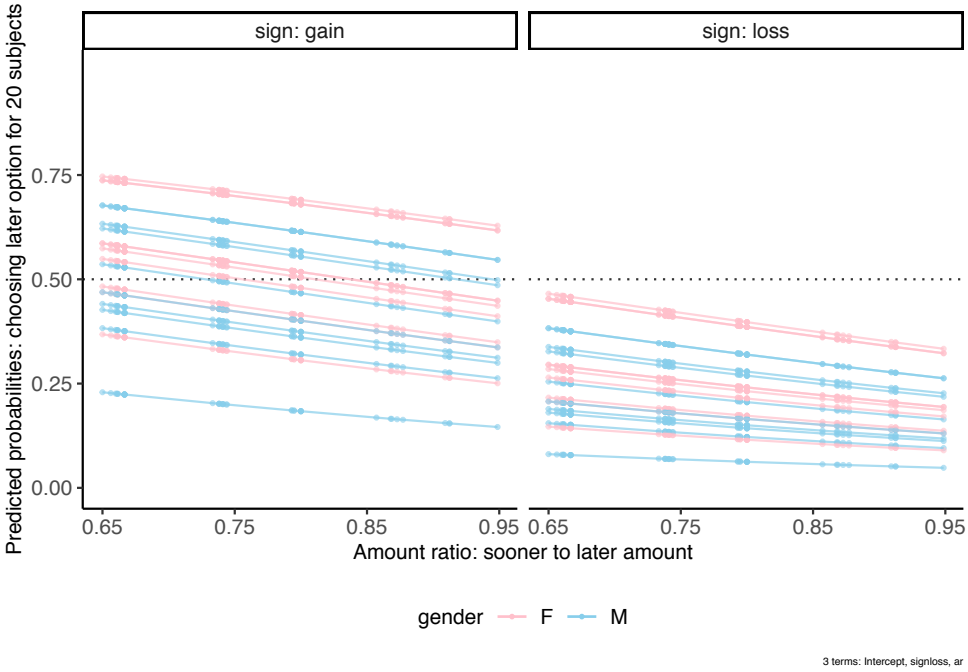


Figure 5.30: Predicted probabilities for model 2.

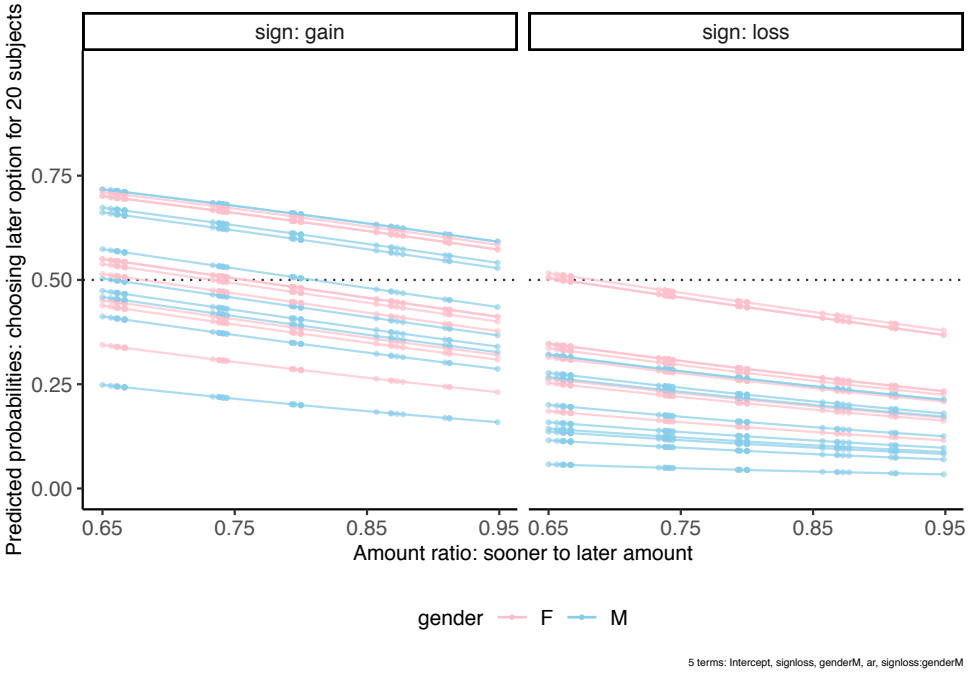


Figure 5.31: Predicted probabilities for model 4.

5.2.4 Diagnostics

The table below presents the diagnostic results for model 4. It provides an additional assessment of the model. The results are separated for gain questions only, loss questions only, and for all questions (both gain and loss).

Table 5.10: Diagnostics results for model 4. The later choice is taken as a "positive", while the sooner choice, a "negative".

Sign	Later choices	Sooner choices	Sensitivity (%)	Specificity (%)	Prevalence (%)	PPV (%)	NPV (%)
all	593	1,007	45.0	86.8	37.1	66.8	72.8
gain	398	402	65.3	69.4	49.8	67.9	66.9
loss	195	605	3.6	98.3	24.4	41.2	76.0

Overall, the model is able to correctly predict the later choice less than half the time (sensitivity: 45%), and the sooner choice most (specificity: 86.8%) of the time. It is able to correctly predict gains about 65–70% of the time. The model correctly predict the sooner loss almost all of the time (specificity). However, the model almost never correctly predicts the later loss (sensitivity). In the predicted probabilities plot for the model, almost all of the predicted probabilities for losses lie below 0.5.

Overall, for the given prevalence of choices, the PPV and NPV indicate that the model correctly predicted later and sooner choices 67% and 73% of the time respectively. For the given prevalence of gains, the model is only able to correctly predict sooner and later gains about two-thirds of the time. For the given prevalence of losses, the model is only able to correctly predict the sooner loss about three-quarters of the time (NPV) but could only correctly predict the later loss less than 50% of the time (PPV).

5.3 Han and Takahashi (2012)

This study had a within and between participant factorial design. The within-participant components are: 2 (sign: gain, loss) × 2 (presentation order: ascending, descending) × 7 (delay: 1 week, 2 weeks, 1 month, 6 months, 1 year, 5 years, 25 years). Order refers to whether the amounts were presented in an ascending or descending order.

There were 50 participants in the study, of whom 12 (24%) were women. No other demographics information were provided in the dataset. Each participant answered 1,148 questions. The sooner amount was always available today and in the Japanese Yen currency (JPY). There were 41 unique values of the sooner amount, ranging from JPY0 to JPY100,000, with the sooner amount increasing by JPY2,500 each time. The later amount was always JPY100,000.

5.3.1 Exploratory data analysis

Figure 5.32 shows the proportion of later choices for each participant coloured by sign. The participants on the x-axis are ordered by increasing proportion of later gains. Each point represents the proportion of later choices from 574 questions.

Just under half (44%) of the proportions are greater than 0.5. The median proportion of later gains and later losses are 0.73 and 0.12 respectively. The proportion of later gains is larger than the proportion of later losses for almost all participants.

Figure 5.33 shows the relationship between the proportion of later choices aggregated over all participants and the time delay till the later option is available, coloured by sign. Each point represents the proportion of later choices from over 4,000 questions.

In Figure 5.33, the relationship between the proportion of later choices and time delay appears to be negative for gains (blue line). The relationship is positive for losses (red line), especially from 6 months onwards. The proportion of later losses is smaller than later gains at each delay until 25 years, where they are about the same.

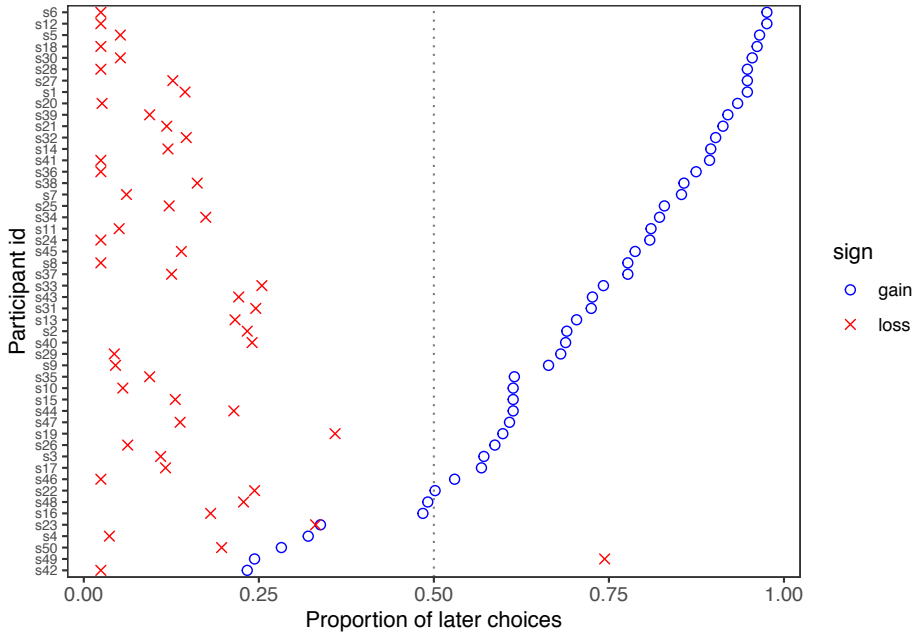


Figure 5.32: Strip chart showing the proportion of later choices for each participant. The participants on the x-axis are ordered by increasing proportion of later gains. The colour and shape of the points represent the two different values of sign: gain and loss.

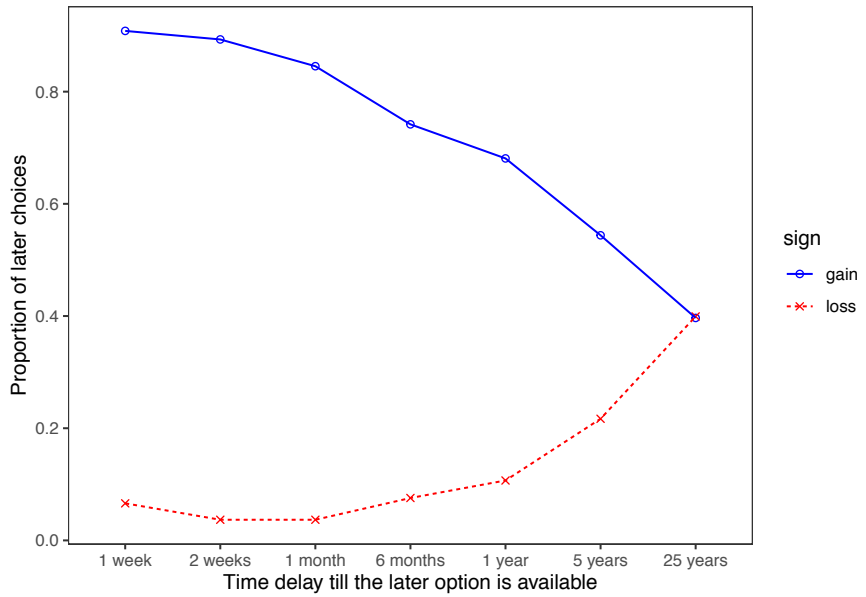


Figure 5.33: Line plot of the proportion of later choices aggregated over all participants and the time delay till the later option is available, coloured by sign.

Figure 5.34 shows the relationship between the proportion of later choices and time delay for each participant. The points and lines are coloured by sign. Participants are ordered by increasing proportion of later gains. There is substantial heterogeneity in choices across participants. There are also many participants who almost always choose the sooner loss.

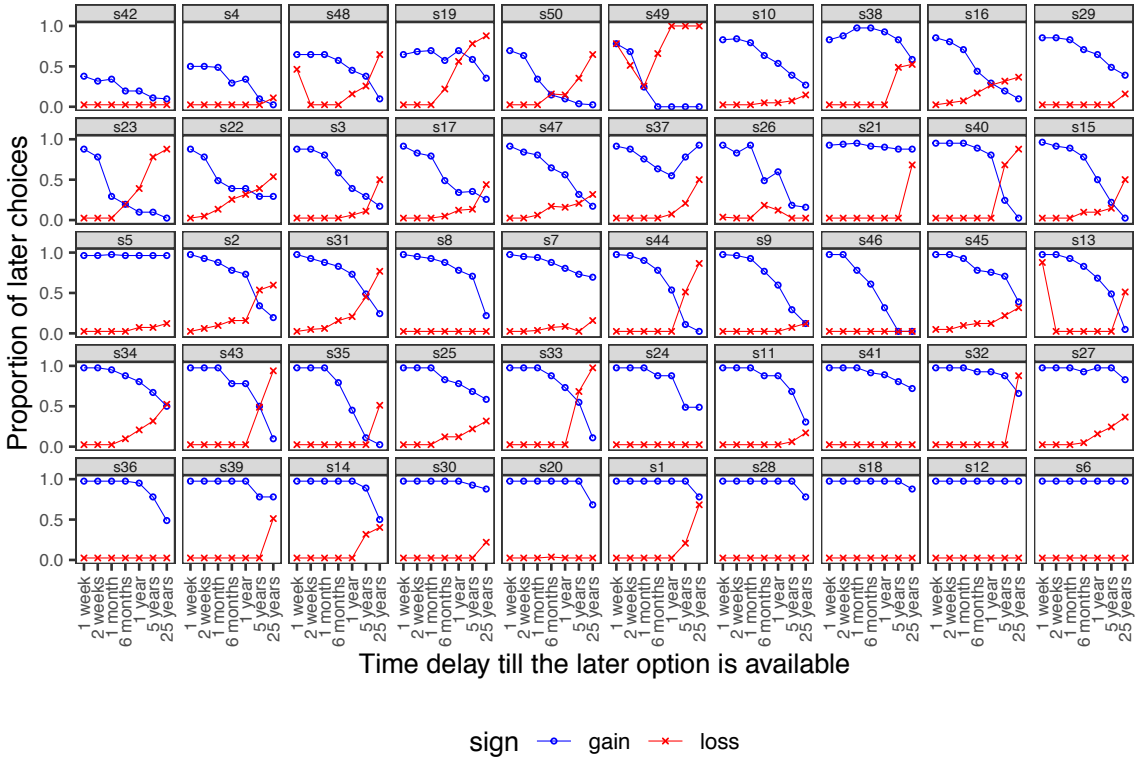


Figure 5.34: Line plots of the proportion of later choices for each participant and time delay coloured by sign. Participants are ordered by increasing proportion of later gains.

Figure 5.35 shows the relationship between the proportion of later choices and amount ratio by gender. The relationship appears positive for gains and negative for losses. As the difference between the sooner and later amounts get smaller, i.e. as the amount ratio increases, the sooner gain and later loss tends to be chosen more often. There is no discernable differences between men and women.

Figure 5.36 shows the relationship between the proportion of later choices and amount ratio for each participant. The lines are coloured by sign. Participants

are ordered by increasing proportion of later gains. There is heterogeneity in the rate of change across participants. However, when the amount ratio is 1, the sooner gain and later loss are almost always chosen while the opposite happens when the amount ratio is zero.

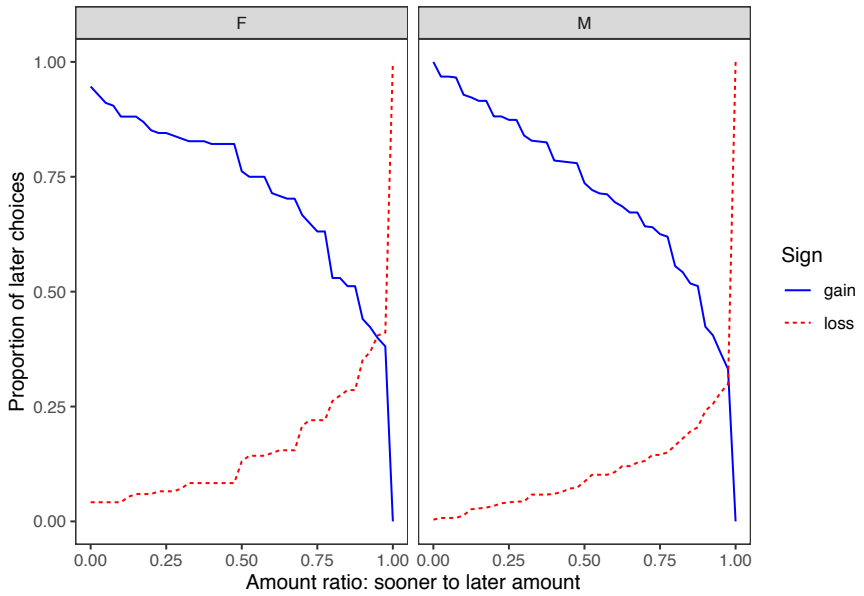


Figure 5.35: Line plots of the relationship between the proportion of later choices and amount ratio by gender. Lines are coloured by sign.

5.3.1.1 Indifference points

Based on the combination of the three within-participant factors, 28 indifference points can be calculated for each participant, half for gains and half for losses. However, there were missing data for 1 participant, who did not have any records of choosing the later option on 41 questions that had the unique factorial combination of: sign: gain; order: descending; and delay: 2 weeks. These missing records will be dropped.

For each unique combination of the within-participant factors, each participant had either a single or no indifference point. An indifference point can not be estimated if participants were always chose either the sooner or later choice. There were 18 (1.3%) indifference points that could not be estimated. Of the 18 instances where an

indifference point could not be estimated, all (50%) were due to participants always choosing the sooner amount. These unestimable indifference points were given the minimum value of the sooner amount (0) if the sooner option was always taken or the maximum value of the sooner amount (100,000) if the later option was always taken.

Figure 5.37 shows box plots of the indifference points (Japanese Yen in thousands) by gender, sign and presentation order of questions. For example, the top left panel compares the distribution of indifference points by presentation order (x-axis) and for women and gains. There is a difference in the median values and spread of values between gains and losses (top vs. bottom panels). However, there are no discernable differences within panels and between panels horizontally. This suggests that the only discernable difference is by sign.

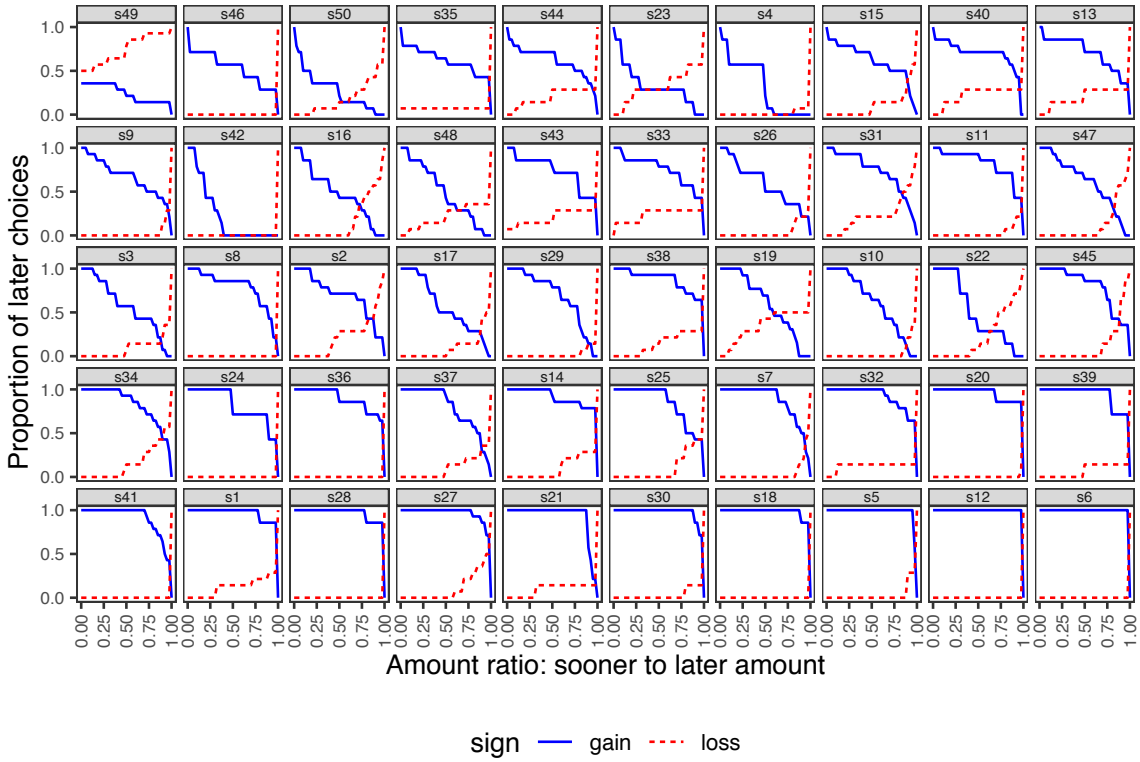


Figure 5.36: Line plots of the proportion of later choices for each participant and amount ratio coloured by sign. Participants are ordered by increasing proportion of later gains.

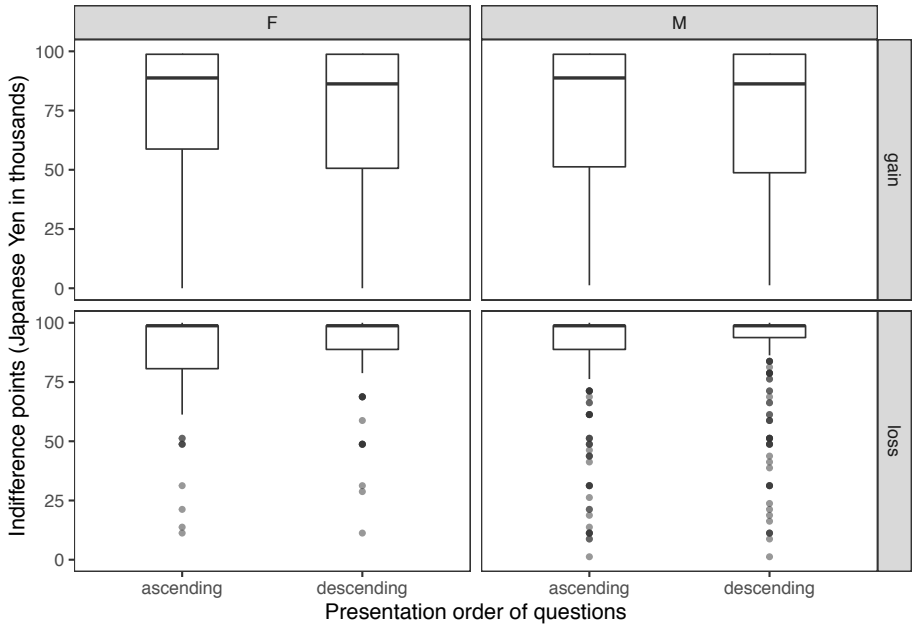


Figure 5.37: Box plots of the indifference points by presentation order, faceted by sign and gender.

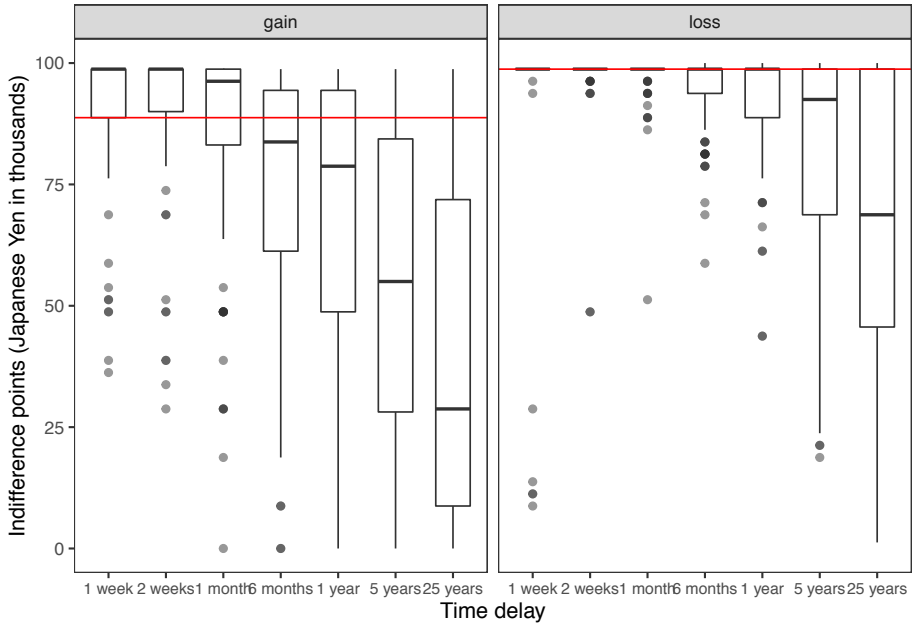


Figure 5.38: Box plots of the indifference points and time delay by sign. Participants are ordered by increasing median values. The horizontal red line represents the overall median value of that group.

Figure 5.38 shows the relationship of the indifference points and time delay by sign. The overall median indifference points (Japanese Yen in thousands) is higher for losses than gains. There is a negative relationship between the indifference points and time delay as the indifference poitns tend to decrease over longer delays. At each time delay, the median value for losses tends to be higher than gains. The heights of the boxes tend to increase as the time delay increases, indicating greater variability.

Figure 5.39 shows the distribution of the indifference points for each participant by sign. There is greater heterogeneity in the indifference points for gains than losses. The median values for losses are closer to the overall median line, partly because many median values are close to 100 on the y-axis. The boxes for losses tend to be shorter than for gains, indicating that indifference points for losses tend to be closer together within a participant.

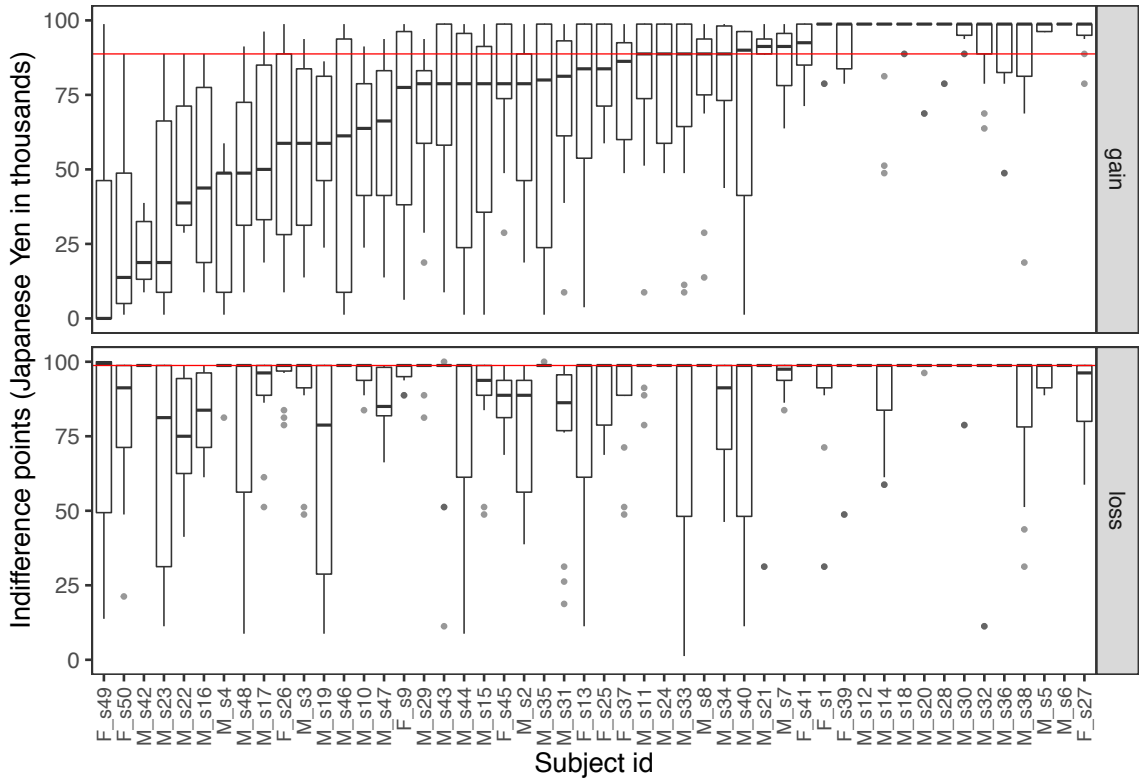


Figure 5.39: Boxplots of the indifference points for each participant by sign. Participants are ordered by increasing median indifference point for gains. The horizontal red line represents the overall median indifference point in that group.

5.3.2 Statistical modelling

There is strong evidence that the between-subject variance is non-zero. The likelihood ratio statistic for testing the null hypothesis that the variance of the average subject is zero, i.e. $\sigma^2_{u0} = 0$, can be calculated by comparing the two-level null model with the correspondong single-level model. The test statistic is very large, 1,836 with 1 degree of freedom from a chi-squared distribution.

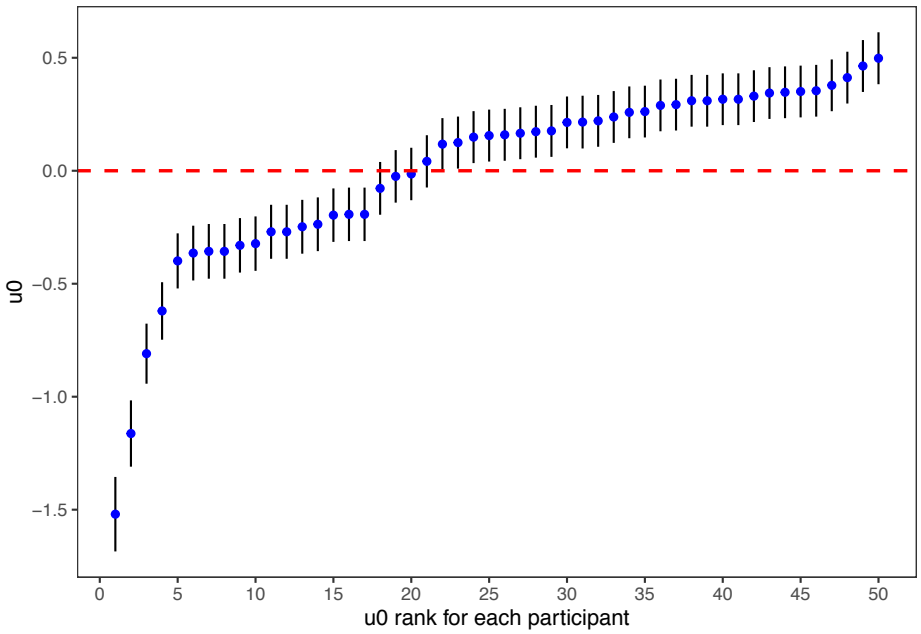


Figure 5.40: Estimated random effects residuals for each participant.

Figure 5.40 shows the estimated residuals for all 50 participants in the sample. For most participants, the 95% confidence interval does not overlap the horizontal line at zero, indicating that the choices for the later option are significantly above or below average (above/below the zero line).

Results from 5 models are displayed in Table 5.11. The models were selected based on the findings from the exploratory data analysis. The estimated coefficients are rounded to two decimal places for presentation purposes. Sign, gender and order are categorical variables. The reference category for sign is gain, for gender is women and for order is ascending.

Table 5.11: Table of multilevel logistic regression results.

	Regression coefficients: Log odds (95% confidence interval)				
Terms	Null model	Model 1	Model 2	Model 3	Model 4
Intercept	-0.32 (-0.44, -0.21)	0.98 (0.81, 1.16)	1.42 (1.25, 1.60)	1.50 (1.32, 1.68)	1.63 (1.27, 1.98)
(sign)loss		-2.97 (-3.01, -2.92)	-3.01 (-3.05, -2.96)	-3.01 (-3.06, -2.96)	-3.01 (-3.06, -2.96)
amount ratio			-0.85 (-0.92, -0.78)	-0.85 (-0.92, -0.78)	-0.85 (-0.93, -0.78)
(order)descending				-0.16 (-0.20, -0.11)	-0.16 (-0.20, -0.11)
(gender)M					-0.16 (-0.57, 0.24)
Random effects					
τ_{00}	0.17	0.38	0.39	0.39	0.38
ICC	0.05	0.10	0.10	0.10	0.10
Participants	50	50	50	50	50
Observations	57,359	57,359	57,359	57,359	57,359
AIC	76,372.8	53,992.1	53,457.0	53,407.7	53,409.1

Reference categories for sign is gain, gender is female and order is ascending.

Across the models, the intraclass correlation coefficient (ICC) ranges between 0.05 and 0.10. This means about 5–10% of the variance is explained by the grouping structure in the population. In Table 5.11, τ_{00} represents the between-subject variance. The between-subject variance is higher in these models compared to models from other studies. Unlike in other studies, the AIC for all models, after the null model, is smaller than the number of observations.

Exploratory data analysis (EDA) findings informed the inclusion of terms in the models. Consistent with the EDA findings, there was no meaningful relationship between gender and choosing the later option. Sign was a strong predictor. Models with time delay and interactions between sign and amount ratio did not converge.

The most substantial reduction in the AIC values of the models were from including sign and amount ratio. For example, model 1 had 2 terms (intercept, sign) and an AIC of 53,992, while model 2 (intercept, sign, ar) had an AIC of 53,457 compared to the null model, which had an AIC of 76,373. The coefficient for the intercept increased substantially when the amount ratio was included.

5.3.3 Predicted probabilities

The plots below show the relationship between the predicted probabilities of choosing the later option and the amount ratio (defined as the sooner amount divided by the larger amount). The relationship is displayed separately in different panels by sign and order. The points and lines are coloured by gender. Only results from selected models are shown.

The relationship between the predicted probabilities and amount ratio is constant for the null model. When the sign term is included, the predicted probabilities shift slightly higher for gains and substantially lower for losses. Once amount ratio is included, the relationship is negative, with some slopes steeper than others. All predicted probabilities are below 0.5 for losses. Models 2, 3 and 4 have very similar patterns, suggesting that including the additional terms after model 2 does not meaningfully change the predicted probabilities. This is consistent with the EDA

findings, which showed no discernable differences between choosing the later amount and presentation order or gender.

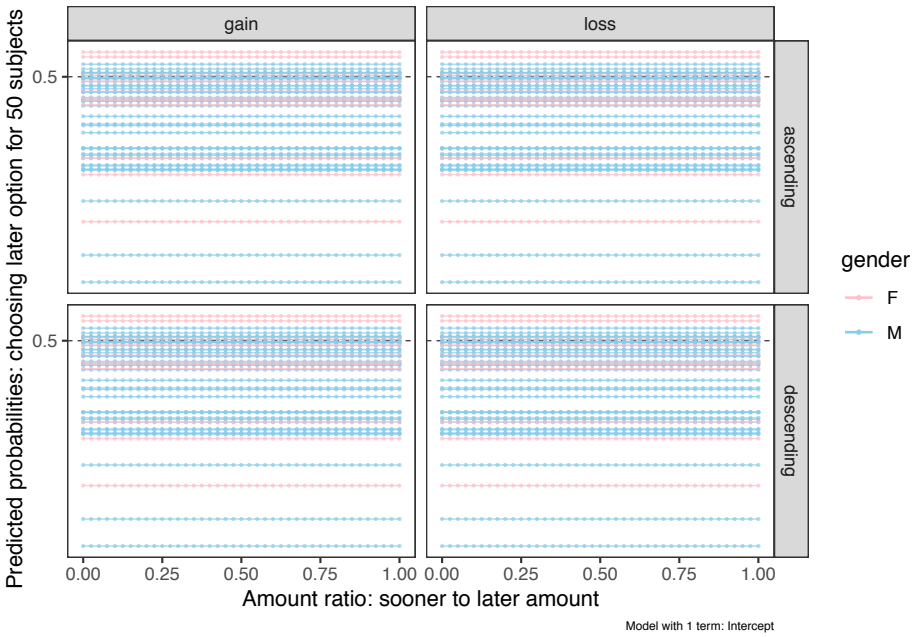


Figure 5.41: Predicted probabilities for null model.

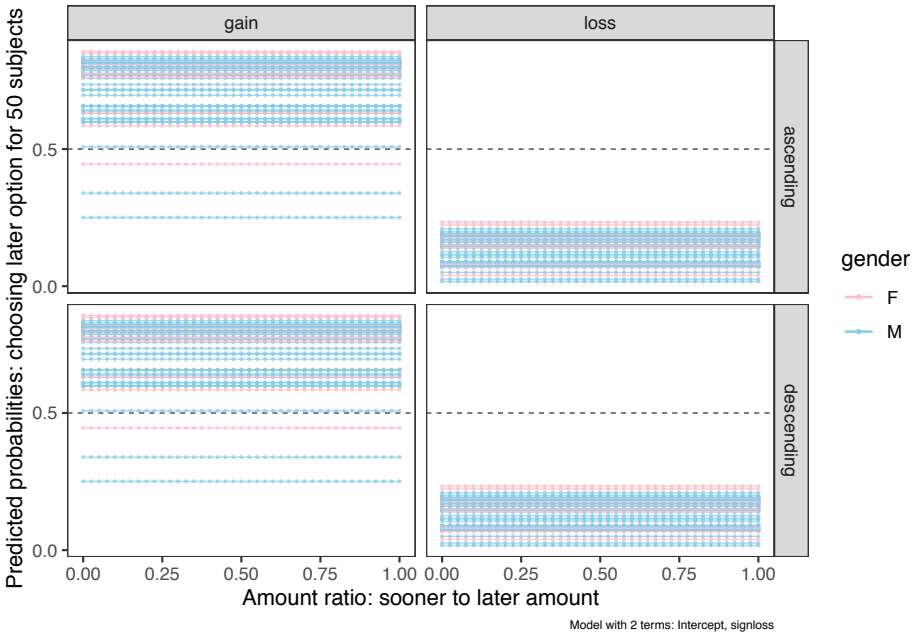


Figure 5.42: Predicted probabilities for model 1.

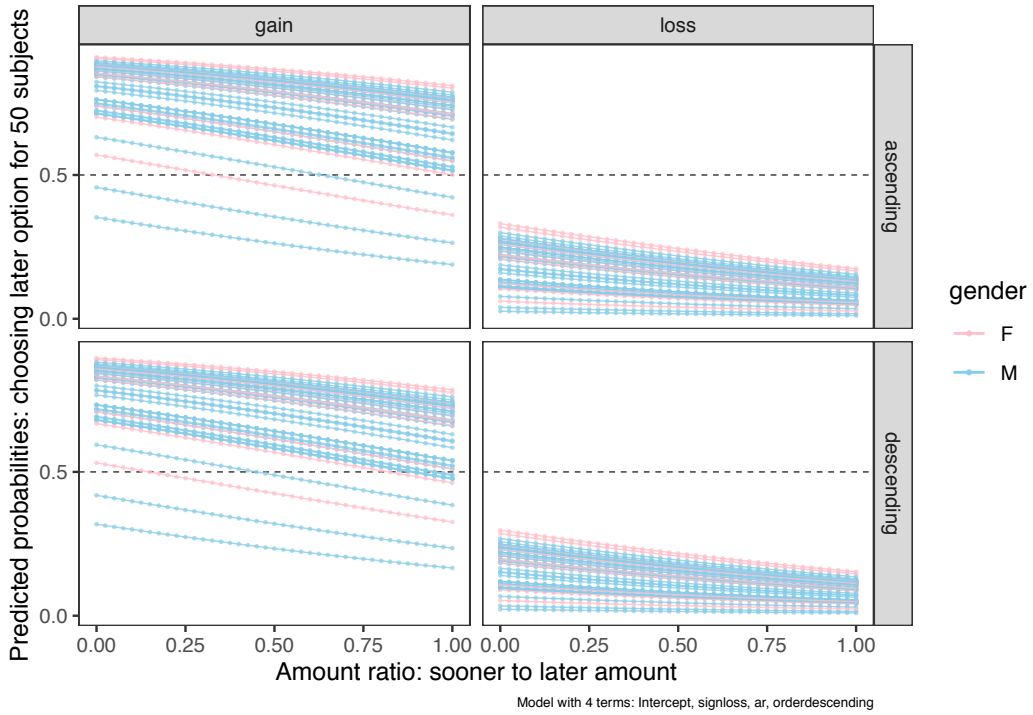


Figure 5.43: Predicted probabilities for model 3.

5.3.4 Diagnostics

Table 5.12 presents the diagnostic results for model 3. It provides an additional assessment of the model. The results are separated for gain questions only, loss questions only, and for all questions (both gain and loss).

Table 5.12: Diagnostics results for model 3. The later choice is taken as a "positive", while the sooner choice, a "negative".

Sign	Later choices	Sooner choices	Sensitivity (%)	Specificity (%)	Prevalence (%)	PPV (%)	NPV (%)
all	24,351	33,008	82.1	79.9	42.5	75.1	85.8
gain	20,504	8,155	97.5	18.8	71.5	75.1	75.0
loss	3,847	24,853	0.0	100.0	13.4	NaN	86.6

Overall, model 3 is able to correctly predict the later choice 82% (sensitivity) and the sooner choice 80% (specificity) of the time. The model is able to correctly predict the

later gain almost all (98%) of the time and the sooner gain only about a fifth of the time. For losses, the model correctly predicts the sooner choice all of the time and the later choice 0% of the time. This can be seen from the predicted probabilities for losses, which are all below 0.5.

For the given prevalence, overall, the model is able to correctly predict the later choice almost 75% of the time (PPV) and the sooner choice 86% of the time (NPV). The model is able to correctly predict the gains three quarters of the time. The model is able to correctly predict the sooner loss close to 90% of the time. It is not able to calculate a PPV for losses as the sensitivity is zero and the specificty is 100%, which makes the denominator zero during the calculations.

5.4 Hardisty et al. (2013)

This section focusses on Experiment 2 of Hardisty et al. (2013). This experiment had a between- and within-participant factorial design. The within-participant factors were: 2 (sign: gain, loss) × 3 (delay in days: 182, 365, 3650). The between-participant factor was the presentation order, which had 2 levels: whether gains were presented first or losses were presented first. There were 53 participants who were presented with gains first and 54 participants who were presented with losses first.

Each participant answered 30 questions on gaining money and 30 questions on losing money. Questions for gains and losses were similar in all respects except for their sign, i.e. whether the monetary amounts were framed as a gain or a loss. The sooner amount was always kept constant at USD300 available today, while the later amount varied from USD250 to USD10,000. There were 3 delays: 6 months, 1 year and 10 years. A participant was asked 10 questions, with varying later amounts, for each delay.

5.4.1 Exploratory data analysis

There were 107 participants (57% women), who were recruited online from Amazon Mechanical Turk. Gender, age or ethnicity were missing for 4 participants. There were also missing data on the choices for gains for one of the 4 participants.

Of the 103 participants with gender, age and ethnicity, 77% reported being 'White', and 10% reported being 'Black or African American', which were similar across genders. The median age for women was 31 years (range: 19 to 56), while the median age for men was 29 years (range: 19 to 62).

Figure 5.44 shows the proportion of later choices for each participant. Each point represents the overall proportion of later choices for one participant. The colour and shape of the points indicate whether the proportion is for gains (blue circle) or losses (red crosses). There is a dotted vertical line at 0.5 on the x-axis to highlight the points to the left and right of the line.

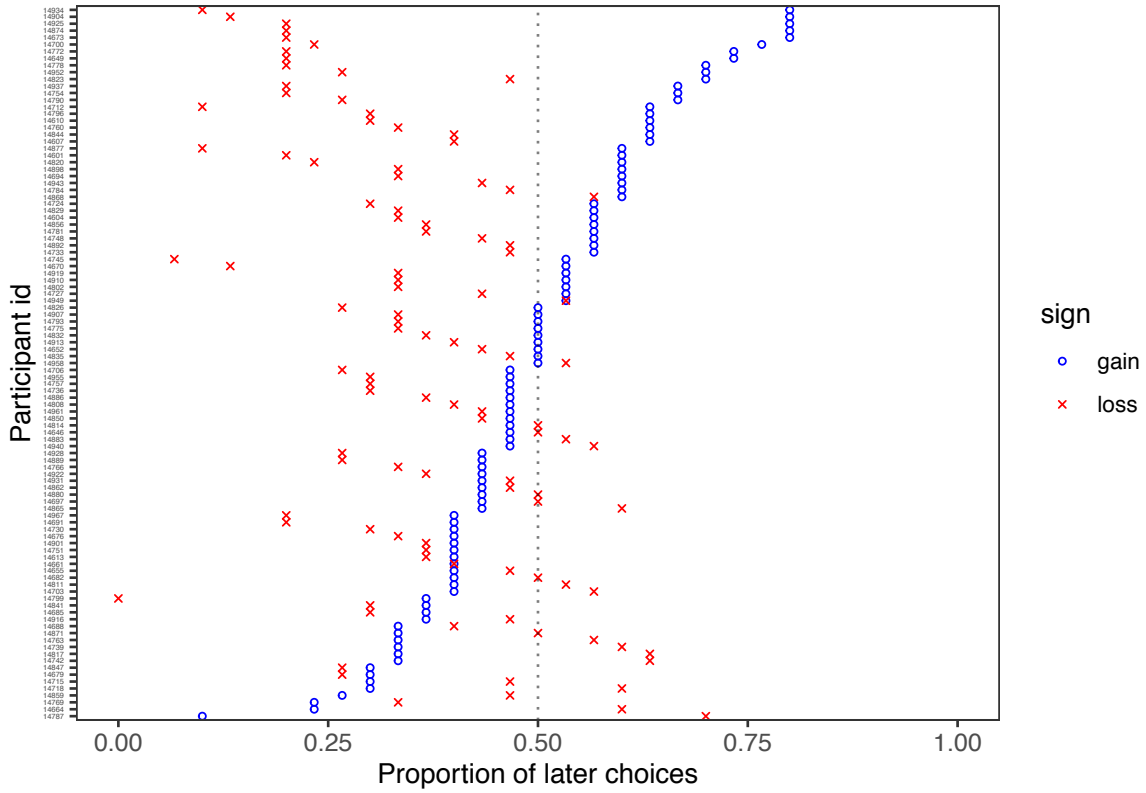


Figure 5.44: Strip chart showing the proportion of later choices for each participant. The participants on the x-axis are ordered by increasing proportion of later gains. The colour and shape of the points represent the two different values of sign: gain and loss.

From Figure 5.44, two-thirds (65%) of the proportions are smaller than 0.5. For losses, 80% of the proportions are smaller than 0.5. For gains, 50% of the proportions are smaller than 0.5. Slightly more than two-thirds (72%) of participants choose the later gain more often than the later loss, which can be seen from the red crosses to the left of blue circles for most participants.

Figure 5.45 shows the distribution of the proportion of later choices by sign and delay, coloured by gender. There does not appear to be meaningful gender differences overall, although the distribution for women is slightly more positively skewed than for losses and a delay of 1 year. The distributions can vary substantially by the delay. For gains, the density of proportions shifts from being negatively to positively skewed as the delay increases. For losses, the distributons are positively skewed for

the delays of half and one year, and slightly more negatively skewed when the delay is 10 years.

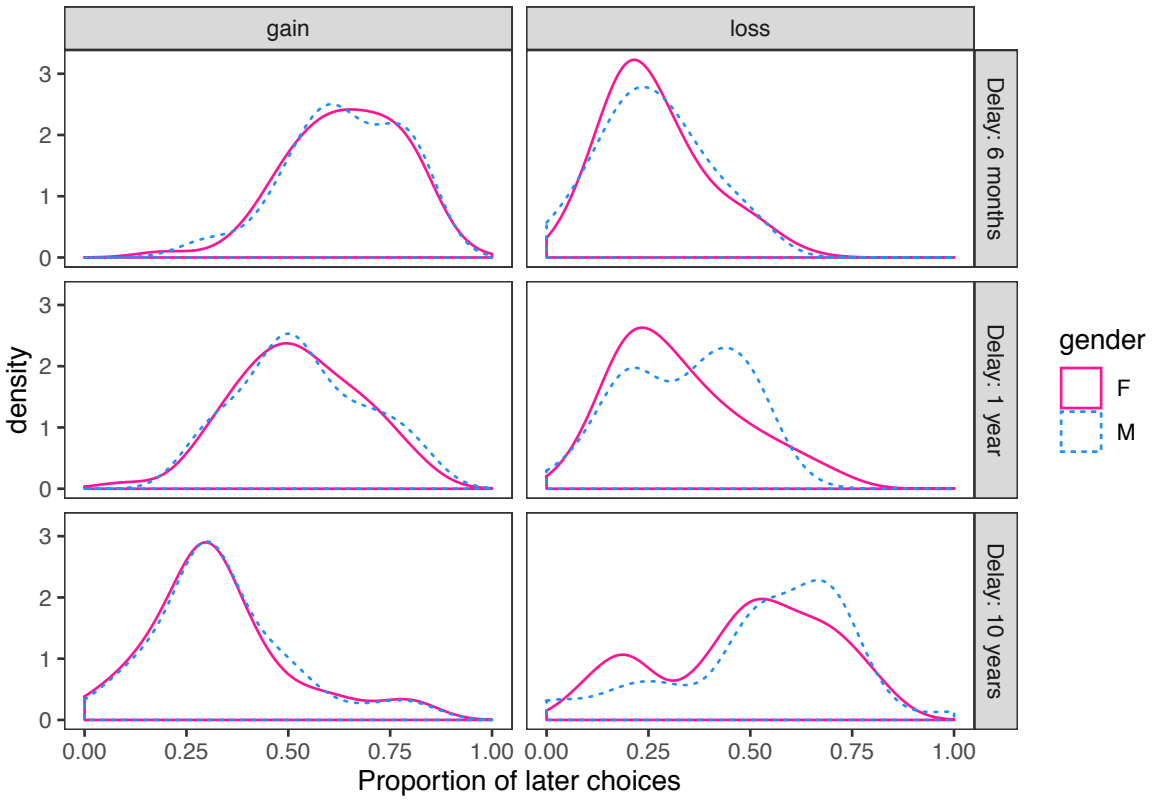


Figure 5.45: Density plots of the proportion of later choices by sign and delay, coloured by gender.

Figure 5.46 shows the relationship between the proportion of later choices and amount ratio by gender. As the amount ratio increases, the proportion of later losses increases while the proportion of later gains decreases rather steeply. The later gain and sooner loss are almost always chosen when the amount ratio is zero while the sooner gain and later loss are almost always chosen when the amount ratio is one. There are no discernable gender differences.

Although not shown, there were no meaningful differences in the relationship between the proportion of later choices and order or age.

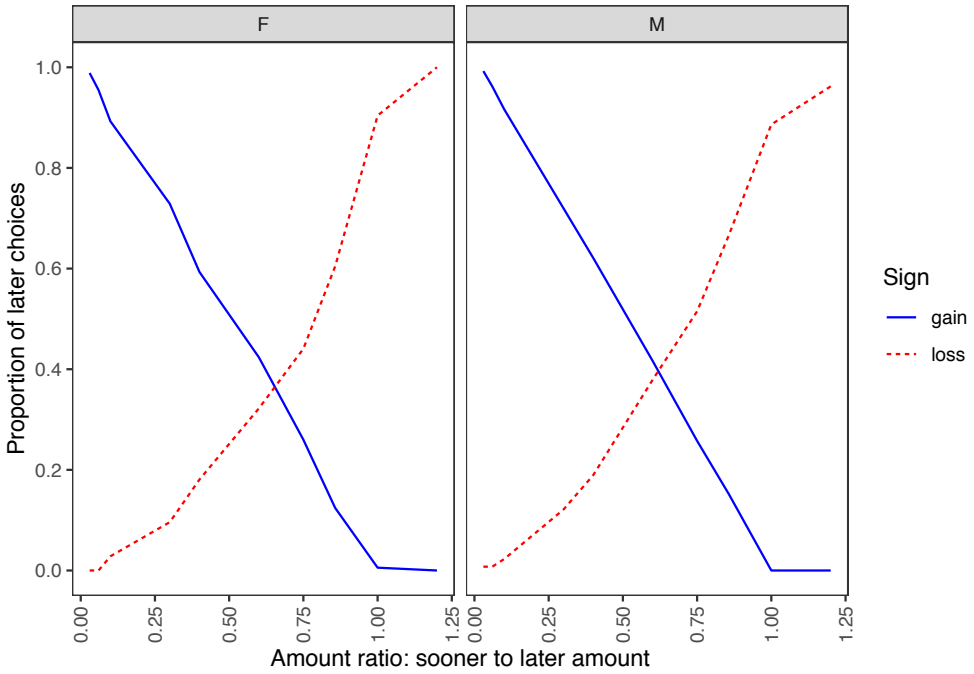


Figure 5.46: Line plots of the relationship between the proportion of later choices and amount ratio by gender. Lines are coloured by sign.

5.4.1.1 Indifference points

Figure 5.47 shows the densities of the ratio of the sooner amount to the indifference points at the different delays, coloured by sign. This ratio, i.e. the sooner amount divided by the indifference point, is similar to the amount ratio: a smaller ratio indicates a larger difference between the sooner amount and indifference point, and vice versa. A ratio of 1 indicates that the sooner amount and indifference point are the same. This ratio was chosen as the sooner amount was constant at USD300, while the later amount varied from USD250 to USD10,000.

Around 2% of the indifference points from the 107 participants could not be estimated accurately because the participants were consistent with their choices, i.e. they always chose either the sooner or later option. These are represented by the short peaks around 0 (the participant always chose the later option) and 1.2 (the participant always chose the sooner option) on the x-axis. As the delay increases, the densities shift from the right side of x-axis to the left side, indicating larger indiffer-

ence points.

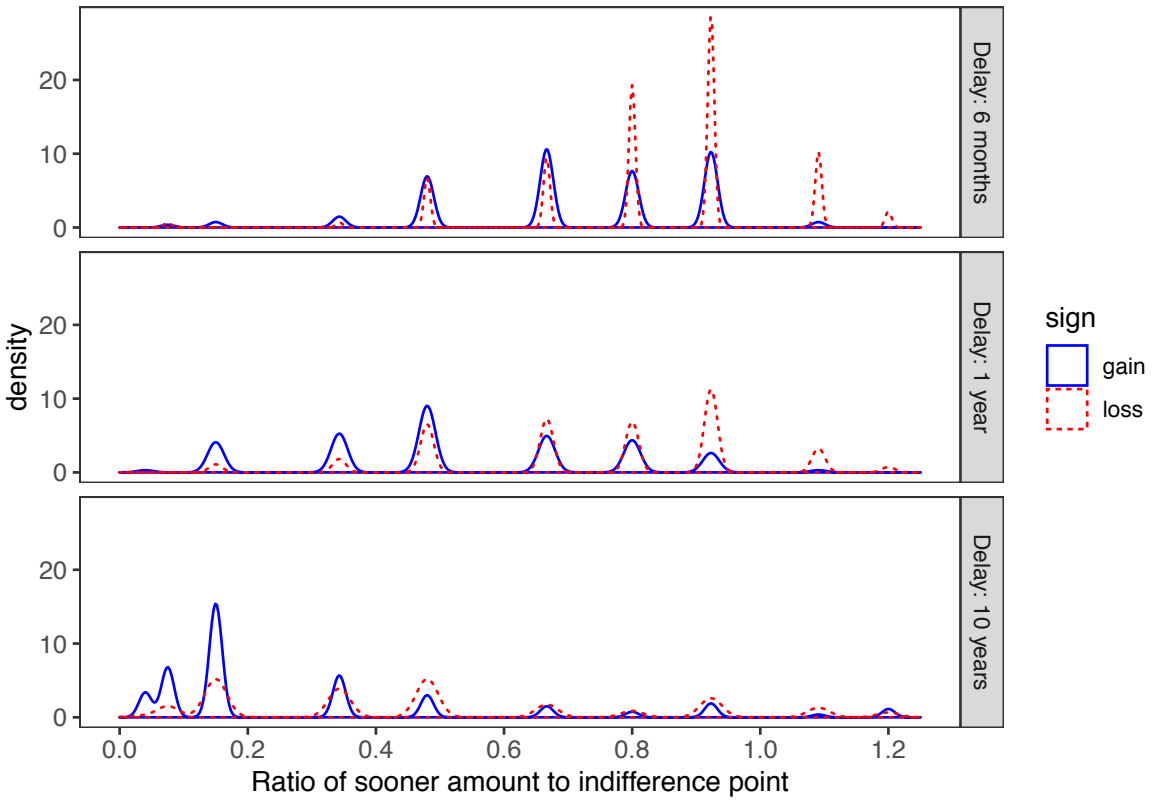


Figure 5.47: Density plots of the ratio of the sooner amount to indifference points at the different delays, coloured by sign.

5.4.2 Statistical modelling

There is an absence of evidence that the between-subject variance is non-zero. The likelihood ratio statistic for testing the null hypothesis that the variance of the average subject is zero, i.e. $\sigma^2_{u0} = 0$, can be calculated by comparing the two-level null model with the correspondong single-level model. The test statistic is 1.9 with 1 degree of freedom from a chi-squared distribution.

Figure 5.48 shows the estimated residuals for all 107 participants in the sample. All participants have 95% confidence intervals that overlap the horizontal line at zero. This suggests that there are no participants with choices for the later option, which are significantly above or below average (above/below the zero line).

None of the models performed particularly well based on the AIC values. Compared to the models from other studies in this chapter, the AIC value did not drop as much when the sign term was added into the model. Adding the interaction between sign and the later delay reduced the AIC the most, relative to the other terms in the preceding models. Although not shown, adding in demographics information, e.g. gender, did not meaningfully improve the results. Models with an interaction between amount ratio and sign produced unreliable results, i.e. very large coefficients.

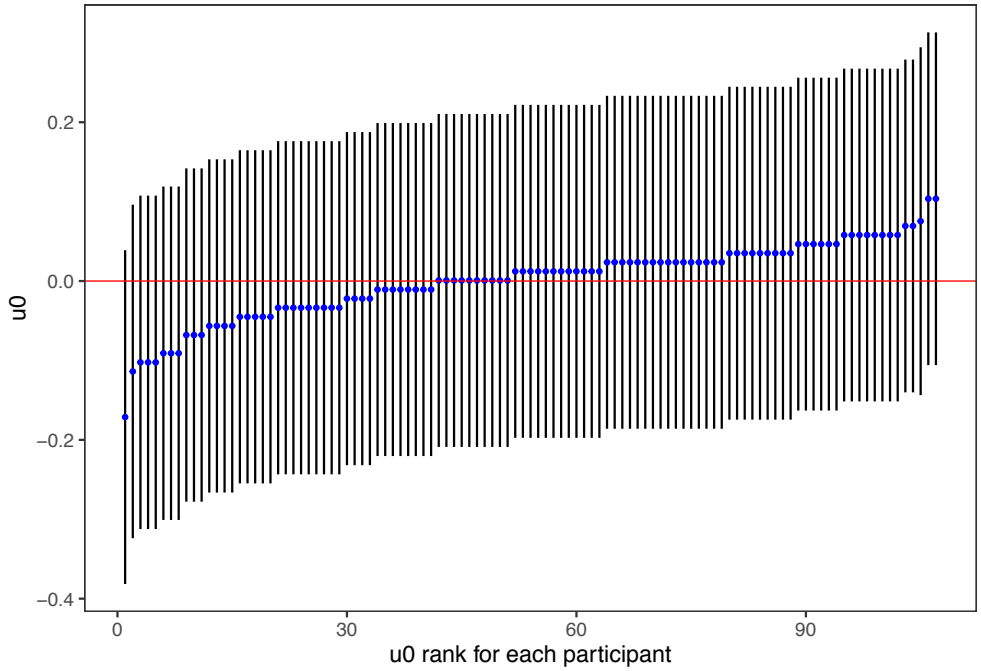


Figure 5.48: Estimated random effects residuals for each participant.

Results from 5 models are displayed in Table 5.13. The models were selected based on the findings from the exploratory data analysis. The estimated coefficients are rounded to two decimal places for presentation purposes. Sign and delay are categorical variables. The reference category for sign is gain and for delay is 6 months.

Across the models, the intraclass correlation coefficient (ICC) ranges between 0 and 0.01. This means about 1% of the variance is explained by the grouping structure in the population. The ICC and τ_{00}, which represents the between-subject variance, is much lower compared to the models from the other studies in this chapter.

Table 5.13: Table of multilevel logistic regression results.

	Regression coefficients: Log odds (95% confidence interval)				
Terms	Null model	Model 1	Model 2	Model 3	Model 4
Intercept	-0.27 (-0.33, -0.22)	0.01 (-0.07, 0.08)	0.07 (-0.03, 0.17)	0.15 (0.03, 0.28)	0.66 (0.51, 0.81)
(sign)loss		-0.57 (-0.67, -0.47)	-0.57 (-0.67, -0.47)	-0.57 (-0.67, -0.47)	-1.62 (-1.81, -1.44)
amount ratio			-0.11 (-0.24, 0.02)	-0.11 (-0.24, 0.02)	-0.12 (-0.25, 0.01)
(tldays)365				-0.11 (-0.24, 0.01)	-0.49 (-0.66, -0.31)
(tldays)3650				-0.14 (-0.27, -0.02)	-1.28 (-1.46, -1.10)
(sign)loss:(tldays)365					0.79 (0.53, 1.04)
(sign)loss:(tldays)3650					2.32 (2.06, 2.57)
Random effects					
τ_{00}	0.01	0.02	0.02	0.02	0.02
ICC	0.00	0.01	0.01	0.01	0.01
Participants	107	107	107	107	107
Observations	6,390	6,390	6,390	6,390	6,390
AIC	8,743.4	8,620.6	8,619.7	8,617.7	8,278.5

Reference categories for sign is gain and delay (tldays) is 6 months.

5.4.3 Predicted probabilities

The predicted probabilities are constant in the null model and adding in the sign term shifts the predicted probabilities for gains higher and losses lower. There is a negative relationship between the predicted probabilities and amount ratio when the amount ratio term is included in the model. Once the interaction term between sign and delay is included in the model, the predicted probabilities for gains decrease as the delay gets longer while the predicted probabilities for losses increase. At the longest delay of 10 years, the predicted probabilities for gains are all below 0.5 while the predicted probabilities for losses are around 0.5.

The predicted probabilities for later losses are always below 0.5, except for the longest delay when the interaction term between sign and delay is included in Model 4. This implies that the models would almost always predict a participant choosing a sooner loss. The predicted probabilities for later gains are around 0.5.

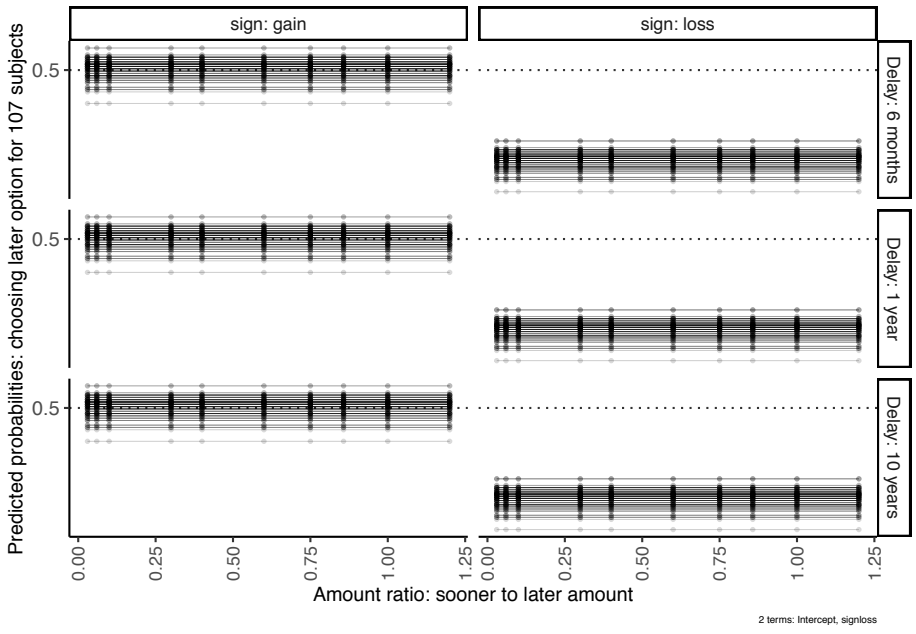


Figure 5.49: Predicted probabilities for model 1.

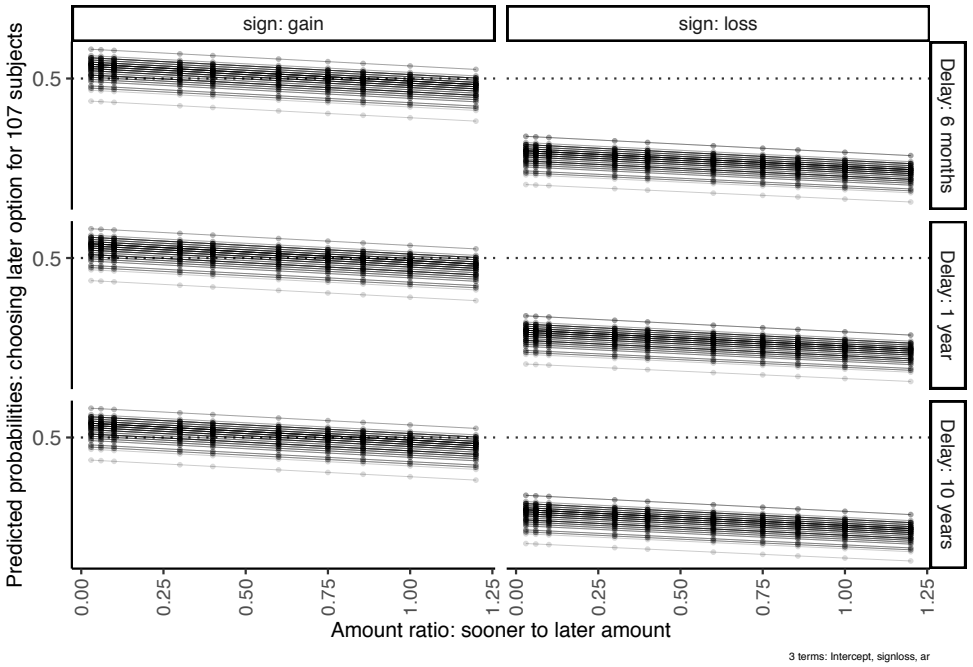


Figure 5.50: Predicted probabilities for model 2.

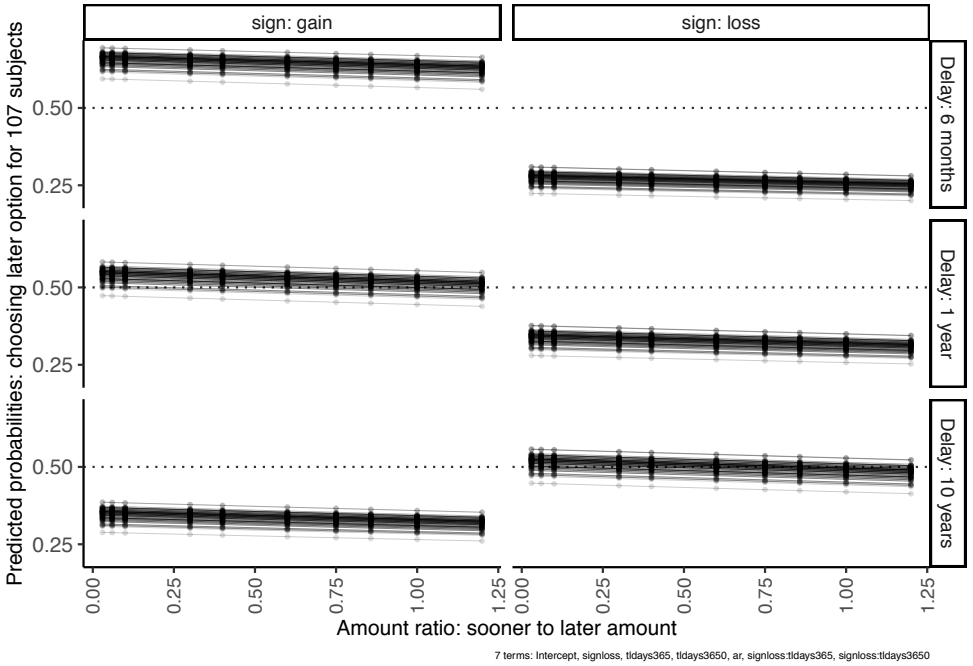


Figure 5.51: Predicted probabilities for model 4.

5.4.4 Diagnostics

Table 5.14 presents the diagnostic results for model 4. It provides an additional assessment of the model. The results are separated for gain questions only, loss questions only, and for all questions (both gain and loss).

Table 5.14: Diagnostics results for model 4. The later choice is taken as a "positive", while the sooner choice, a "negative".

Sign	Later choices	Sooner choices	Sensitivity (%)	Specificity (%)	Prevalence (%)	PPV (%)	NPV (%)
all	2,763	3,627	52.9	69.2	43.2	56.7	65.9
gain	1,595	1,585	76.6	51.9	50.2	61.6	68.8
loss	1,168	2,042	20.5	82.7	36.4	40.4	64.5

Overall, model 4 is able to correctly predict the later choice about half (53%) of the time (sensitivity) and the sooner choice about 70% of the time (specificity). The model is able to correctly predict later gains 77% of the time and the sooner gain 52% of the time. For losses, it is able to correctly predict the later choice a fifth (21%) of the time and the sooner choice 83% of the time.

For the given prevalence, overall, the model is able to correctly predict the sooner (NPV) and later (PPV) choices about 57% and 66% of the time respectively. The model is able to correctly predict the sooner and later gains about 69% and 62% of the time respectively. For losses, it is able to correctly predict the sooner choice 65% of the time and the later choice 40% of the time.

5.5 Hardisty and Weber (2009) Experiment 1

The dataset consists of 2 studies and 183 participants in total. There were 65 participants in Study 1 and 118 participants in Study 2. In both studies, each participant answered 20 questions, half of which were for gains and the other half for losses. In both studies, participants were recruited online.

In both studies, the questions always had the sooner option available today and the later option available in a year. The sooner amount was always kept constant at USD250, while the later amount varied from USD230 to USD410, increasing by USD20 each time.

There were 100 (2.7%) missing values out of the 3,660 observations. Personal characteristics, e.g. gender and age, were missing for 2 participants.

5.5.1 Exploratory data analysis

This section will focus on Study 1. Participants were recruited online via Amazon Mechanical Turk. There were 42 females (65%), 22 males (34%) and 1 participant with missing information for all personal characteristics. For the purposes of the exploratory analysis, the 1 participant with missing information will be dropped.

Participants had either zero or one switching point because 16 (18%) participants were excluded on the basis that they 'switched back and forth more than once, or switched in a manner that would make sense only if they preferred more losses or fewer gains (i.e., preferring $150 now to $250 in 1 year yet also preferring $230 in 1 year to $150 now)' (Hardisty and Weber 2009, 331).

5.5.1.1 Participants' characteristics

Table 5.15 summarises the characteristics of participants. All characteristics, including gender, are self-reported. Percentages in tables are rounded to the nearest whole number. Gender differences in self-reported characteristics will be highlighted.

Only a third of women do not have children compared to three-quarters of men. More than half the men report being single while more than half the women report being married. Most men report being students or workers/farmers (72%) while women have a variety of roles. Women tend to be a few years older than men.

Table 5.15: Characteristics of participants by gender.

	Female	Male
Number of participants	42	22
Median age (range)	31.5 years (18–49)	24 years (16–45)
Eduation (%)		
Bachelor's degree and above	48	41
Below Bachelor's degree	52	59
Number of children (%)		
None	33	73
1 or 2	52	18
3 or more	14	10
Marital status (%)		
Single	24	55
Married	62	32
Others	15	14
Annual household income (%)		
Less than $50,000	56	46
$50,000 to $99,999	31	50
$100,000 or more	14	4
Job type (%)		
None/household	26	0
Student	19	45
Worker/farmer	0	27
Employee/manager	20	28
Others	26	0

Percentages rounded to nearest whole number, with some rounding error.

5.5.2 Proportion of later choices

Figure 5.52 shows the proportion of later choices for each participant. Each point represents the overall proportion of later choices for one participant. The colour and shape of the points indicate whether the proportion is for gains (blue circle) or losses (red crosses). There is a dotted vertical line at 0.5 on the x-axis to highlight the points to the left and right of the line.

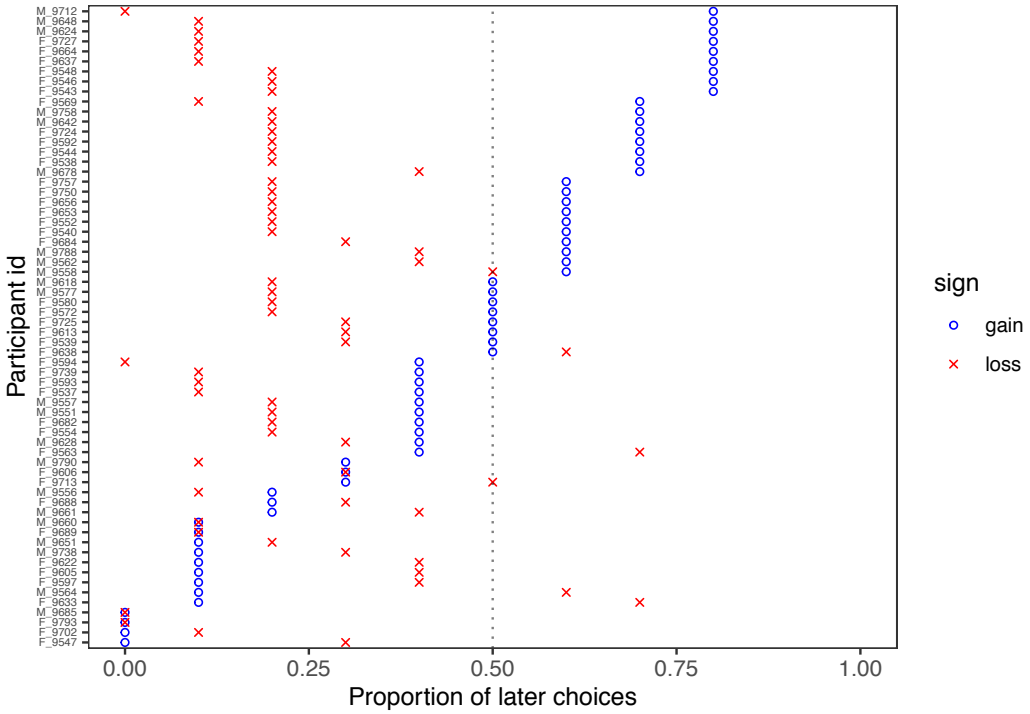


Figure 5.52: Strip chart showing the proportion of later choices for each participant. The participants on the x-axis are ordered by increasing proportion of later gains. The colour and shape of the points represent the two different values of sign: gain and loss.

From Figure 5.52, two-thirds (68%) of proportions are smaller than 0.5, which implies that most choices are for the sooner option. For losses, almost all (91%) of the points

are smaller than 0.5. For gains, just under half (42%) the points are greater than 0.5. Most (70%) participants choose the later gain more often than the later loss.

Figure 5.53 shows the box plots of the proportion of later choices for gender by sign. The median proportion of later losses is lower than that of gains (0.2 vs. 0.5). Men have a slightly lower median proportion of later gains than women. There are no discernable gender differences for the proportion of later losses. The boxes for men are taller, indicating greater variability in choices.

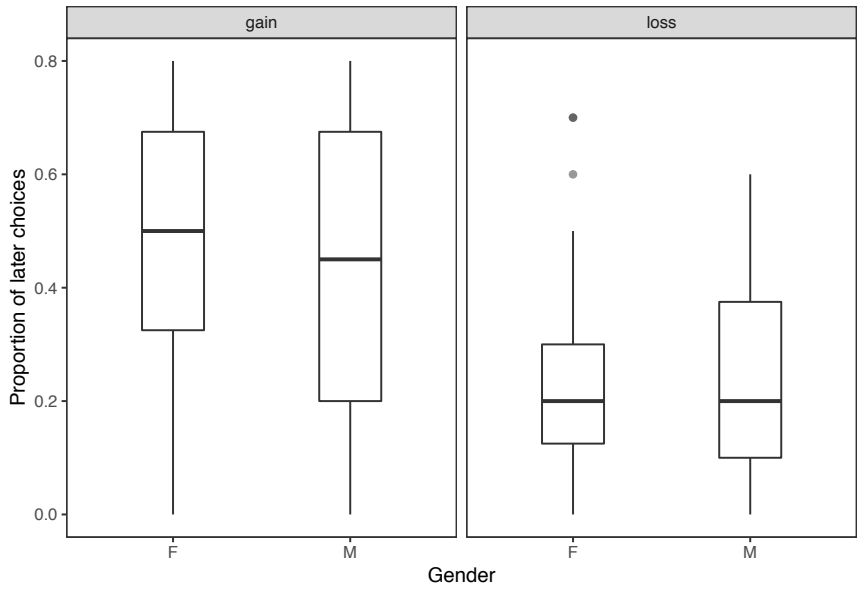


Figure 5.53: Box plots of the proportion of later choices for gender by sign.

The jittered strip charts below show the relationships between the proportion of later choices and marital status, education, and job type. The relationship is shown separately for gains and losses. Points have different colours and shapes based on gender (pink circles for women and blue crosses for men). Each point represents the proportion of later choices from one participant. The horizontal lines represent the median value.

The proportion of later gains tended to be higher than the proportion of later losses. There tended to be greater differences between the different categories for gains than losses.

Figure 5.54 shows the relationship between the proportion of later choices and marital status. On the x-axis, the marital status categories are ordered by increasing median value.

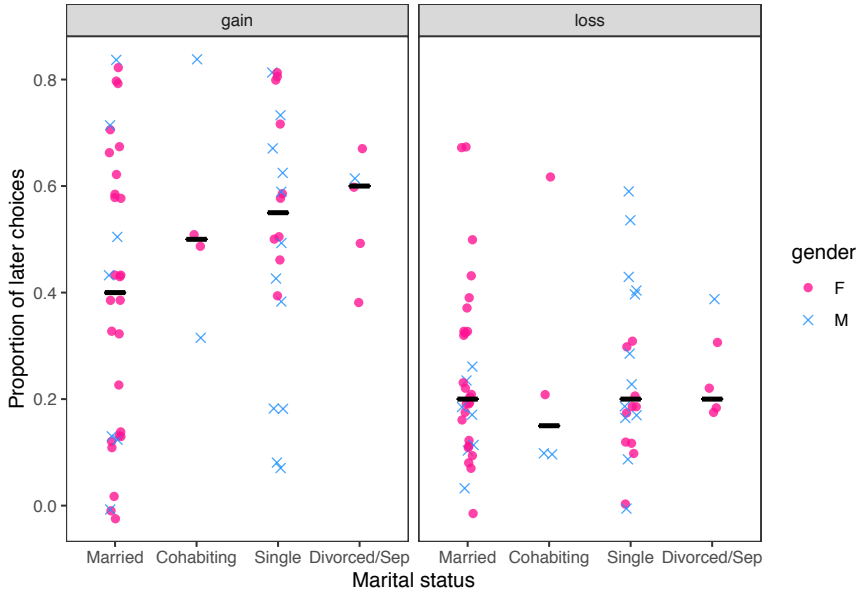


Figure 5.54: Jittered strip charts of the relationship between the proportion of later choices and marital status by sign. Points are coloured by gender. Marital status categories are ordered by increasing median value. Horizontal black line represents the median value of that group.

Figure 5.55 shows the relationship between the proportion of later choices and degree status. Several categories with only a few data points were combined, e.g. 'Others' contain 'No' and 'Professional' degrees, and 'Associate' combines the academic and occupational associate degrees.

Figure 5.56 shows the relationship between the proportion of later choices and job type. The category 'Entrepreneur' was combined with the category, 'Other'. On the x-axis, the job type categories are ordered by increasing median value.

Figure 5.57 shows the relationship between the proportion of later choices and age split by sign and gender. For example, the top left panel represents the relationship for gains and female. There is no apparent relationship between proportion of later choices and age.

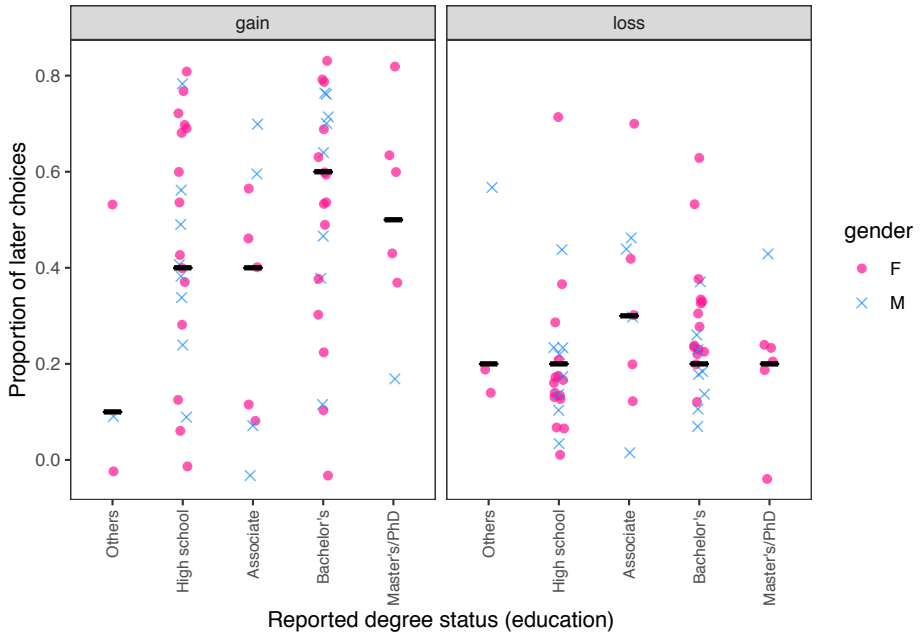


Figure 5.55: Jittered strip charts of the relationship between the proportion of later choices and degree status by sign. Points are coloured by gender. Horizontal black line represents the median value of that group.

Figure 5.58 shows the relationship between the proportion of later choices and amount ratio for men and women. Both men and women have similar patterns for gains and losses. For gains, as the amount ratio increases, the proportion choosing the later gain decreases and reaches 0 when the amount ratio is 1. For losses, the proportion of choosing the later loss increases and reaches close to 1 when the amount ratio is 1.

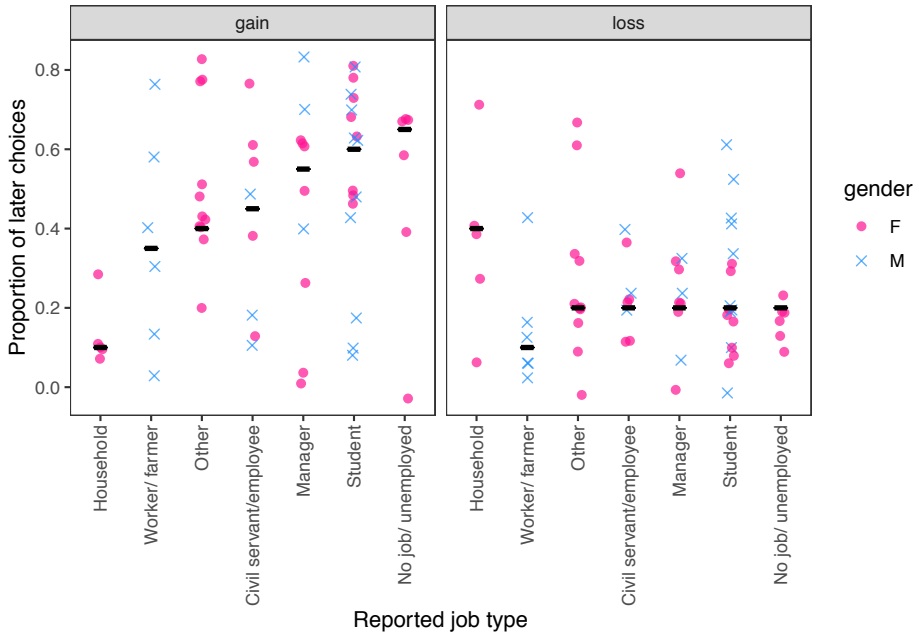


Figure 5.56: Jittered strip charts of the relationship between the proportion of later choices and job type by sign. Points are coloured by gender. Job type categories are ordered by increasing median value. Horizontal black line represents the median value of that group.

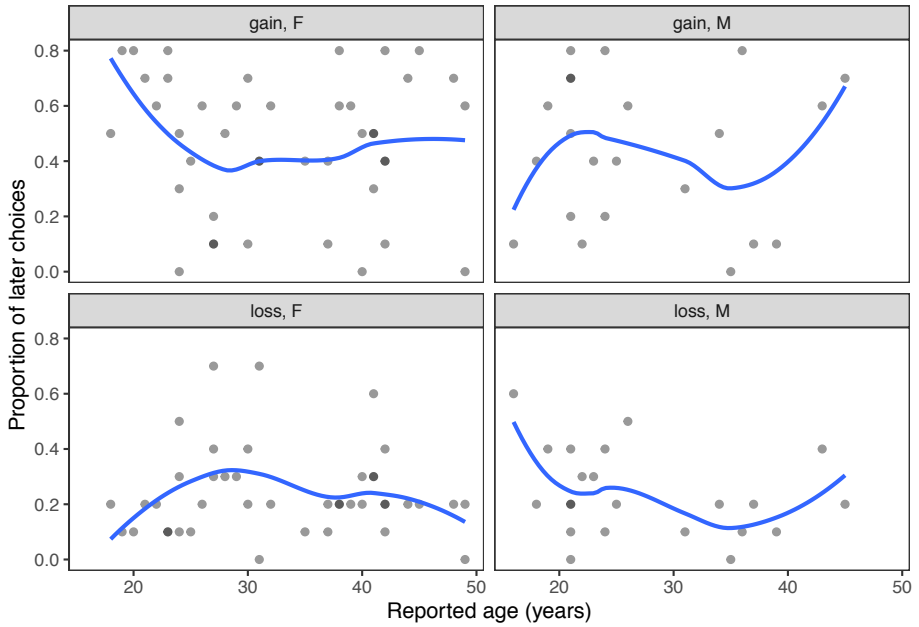


Figure 5.57: Scatter plots of the relationship between the proportion of later choices and age by gender and sign.

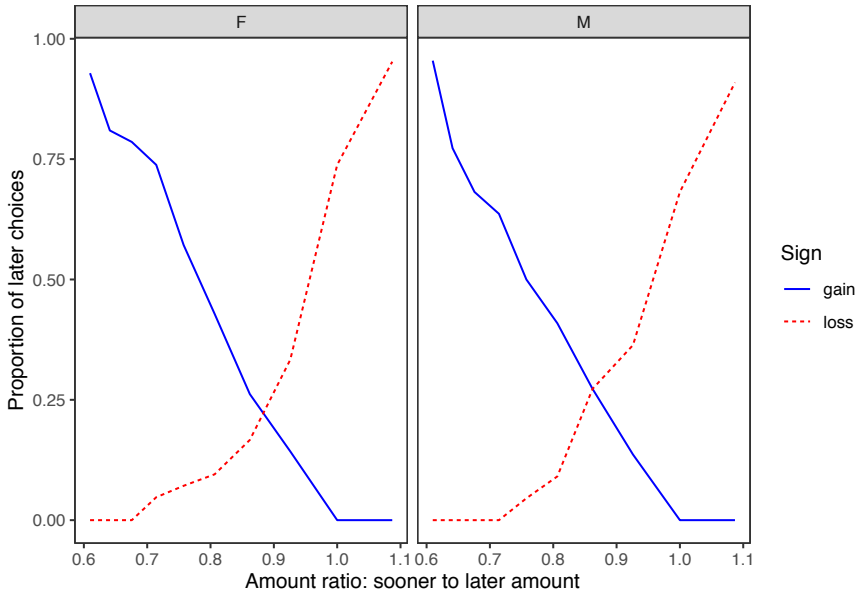


Figure 5.58: Line plots of the relationship between the proportion of later choices and amount ratio by gender. Lines are coloured by sign.

5.5.2.1 Indifference points

There were 6 participants (9% of all 64 participants) who always chose the sooner option and thus did not have a switching point, of which 2 participants did not switch for both gains and losses.

Figure 5.59 shows each choice each participant made for gains and losses. The participants are ordered by a decreasing number of later choices for gains. When given a choice between gaining USD250 now or gaining USD230 in 1 year, every participant chose the sooner gain. Conversely, when given a choice between losing USD250 now or losing USD230 in 1 year, almost every participant chose the later loss.

Figure 5.60 shows the densities of the indifference points for participants coloured by sign. The indifference point is calculated by adding the sooner amount where participants first switched preferences (e.g. from preferring the later to the sooner option) and the sooner amount before the first switch and then dividing this by 2. Participants who did not switch and always chose the sooner amount were given a

value of 230, which is the smallest sooner amount.

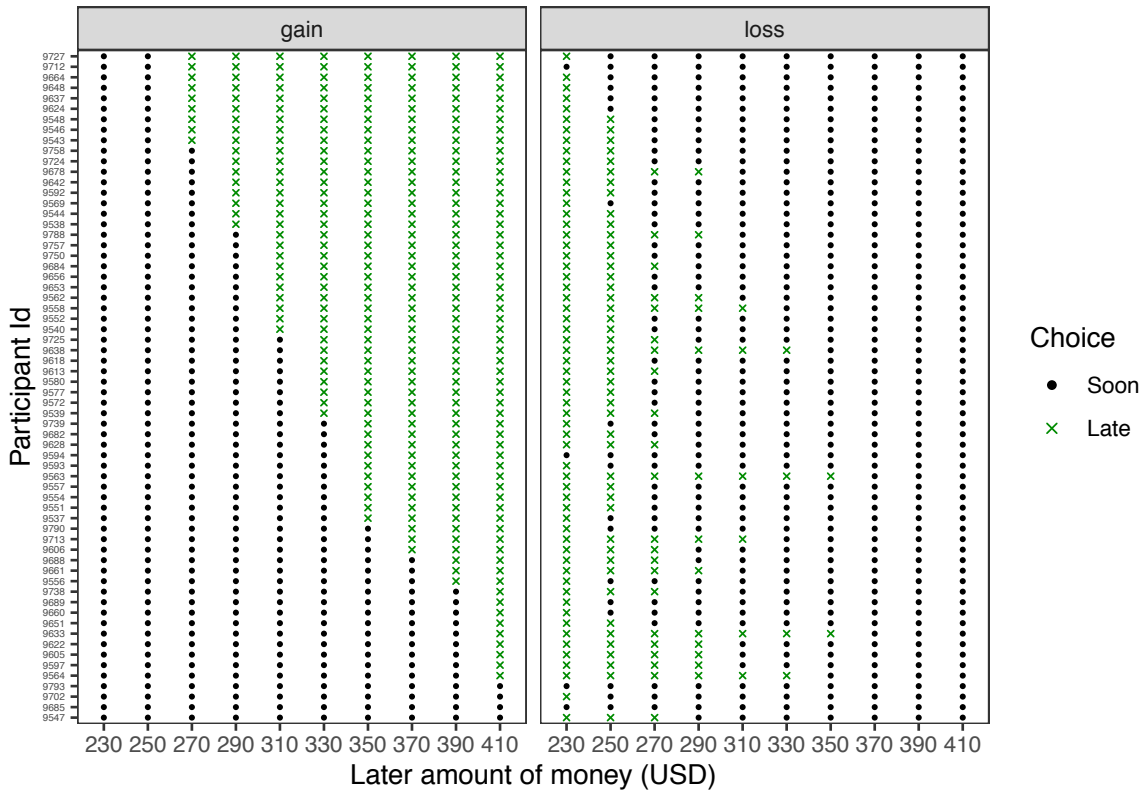


Figure 5.59: Strip charts of each choice each participant made by sign. Participants are ordered by decreasing number of later choices for gains.

For losses, the peaks tend to occur at the smallest values (e.g. 240, 260, 280), with no peaks at 380 and 400. For gains, the peaks tend to be uniform between the values 260 and 340 and at 400, with no peak at 240. This suggests that participants tend to switch preferences earlier for losses than gains. The 6 participants who did not switch at all are reflected by the peaks at 230 on the x-axis.

Figure 5.61 shows the difference in indifference points for each participant. The difference for each participant was calculated by subtracting the indifference point for losses from the indifference point for gains. If there were no difference, the points would lie on the zero line. If the indifference points for gains were higher than losses, then the points would lie above the zero line. If the indifference points for losses were higher than losses, then the points would lie below the zero line.

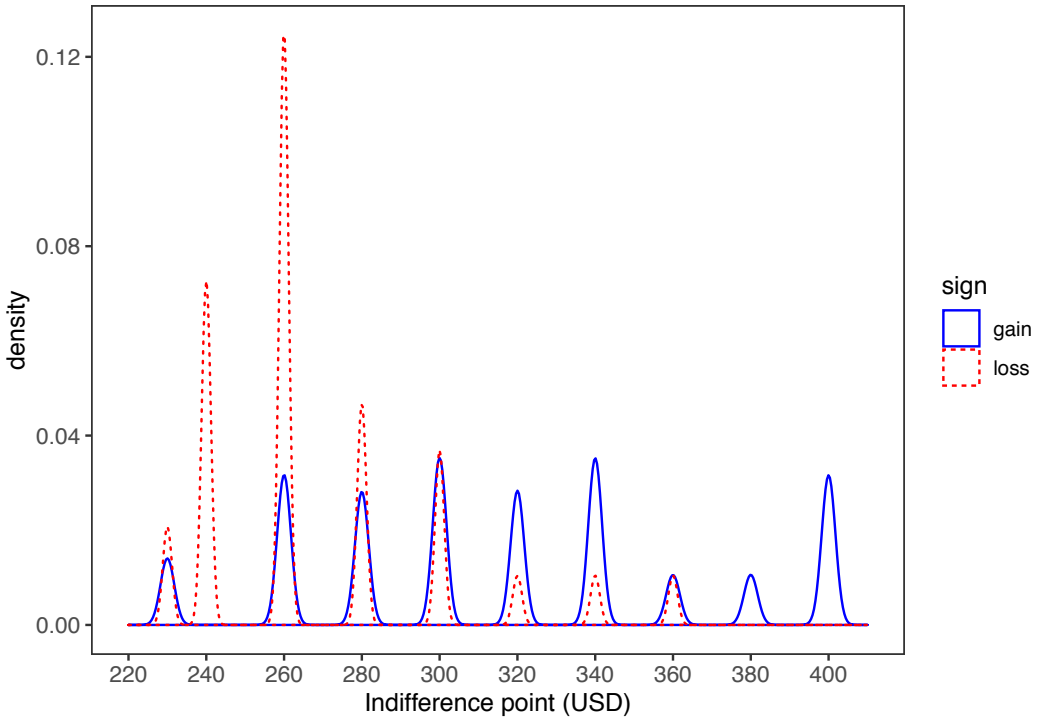


Figure 5.60: Density plots of the indifference points coloured by sign.

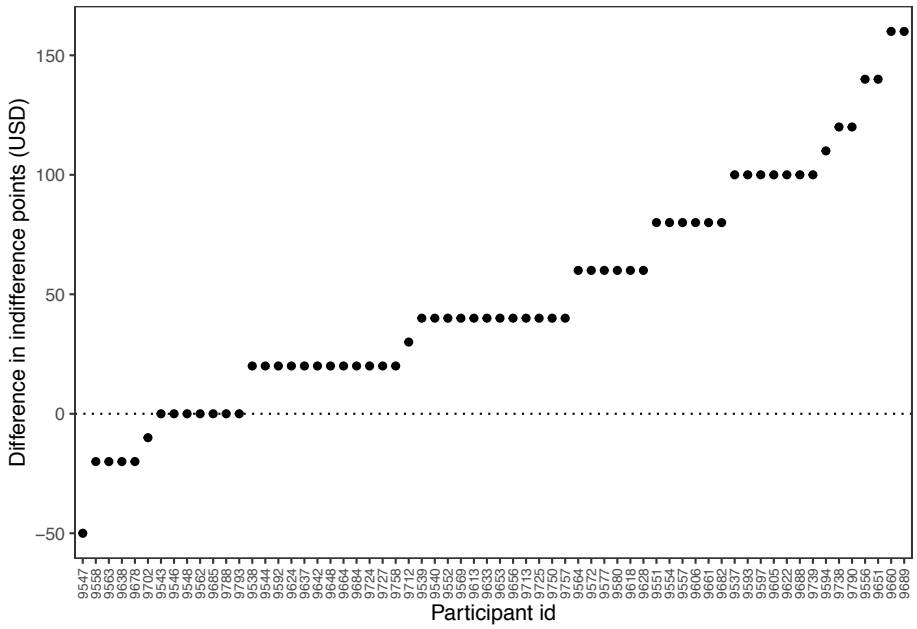


Figure 5.61: Strip chart of the difference in indifference points for gains and loss for each participant, with zero line of no difference.

Figure 5.61 shows that most points lie above the zero line, which means that most participants had a higher indifference point for gains than losses. A few participants had the same indifference point for gains and losses. The points that lie below the zero line shows that there were a few participants who had a higher indifference point for losses than gains.

5.5.3 Statistical modelling

There is evidence that the between-subject variance is non-zero. The likelihood ratio statistic for testing the null hypothesis that the variance of the average subject is zero, i.e. $\sigma^2_{u0} = 0$, can be calculated by comparing the two-level null model with the correspondong single-level model. The test statistic is 8.9 with 1 degree of freedom from a chi-squared distribution.

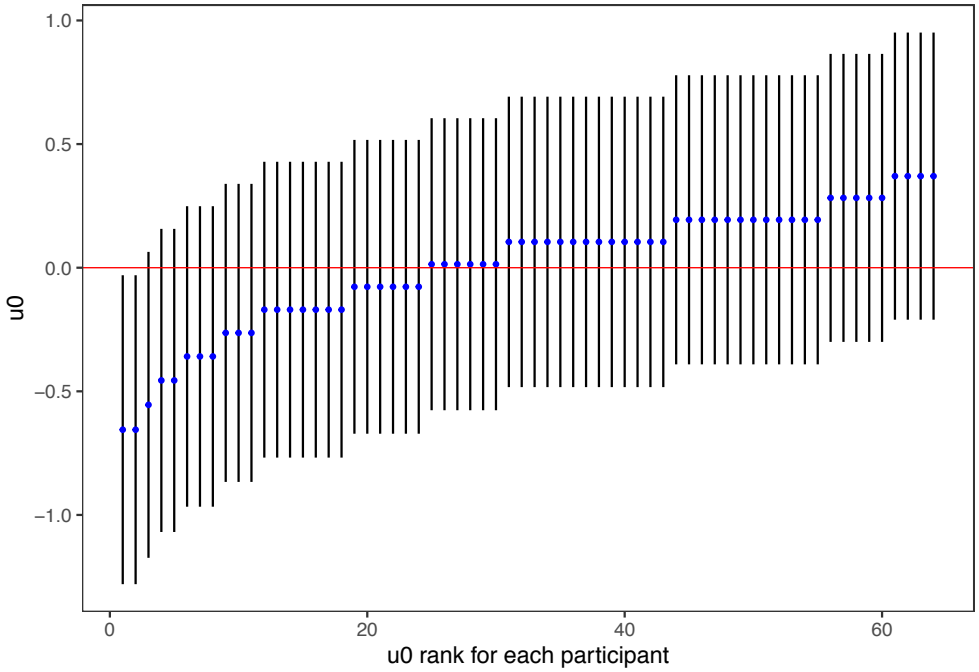


Figure 5.62: Estimated random effects residuals for each participant.

Figure 5.62 shows the estimated residuals for all 64 participants in the sample. Only two participants have a 95% confidence interval that does not overlap the horizontal

line at zero. This indicates that the choices for the later option are significantly above or below average for only a few participants (above/below the zero line).

Results from 5 models are displayed in Table 5.16. The models were selected based on the findings from the exploratory data analysis. The estimated coefficients are rounded to two decimal places for presentation purposes.

Sign, gender and education and marital status are categorical variables. Marital status was reduced by combining the categories of those who live together and those who are divorced or separated. Degree status was reduced from 8 categories to 2 categories: No bachelors, Bachelors and above. The latter category contains participants with a Bachelor's, Master's or Doctoral degree. The reference category for sign is gain, for gender is women, for education is education diplomas before a Bachelor's degree, and for marital status is single.

Across the models, the intraclass correlation coefficient (ICC) ranges between 0.04 and 0.05. This means about 4–5% of the variance is explained by the grouping structure in the population. In Table 5.16, τ_{00} represents the between-subject variance.

None of the models performed particularly well based on the AIC values. The model with only the intercept and sign performed well relative to the other models with an AIC of 1,581. Adding in demographics information such as gender, education and marital status only improved the model slightly. Models with an interaction between amount ratio and sign produced unreliable results, i.e. very large coefficients.

5.5.4 Predicted probabilities

The predicted probabilities are constant when the amount ratio term is not included. Once the amount ratio is included, the relationship between the predicted probabilities and later amount becomes slightly positive. The predicted probabilities for later losses are well below 0.5, which implies that the models almost always predict a participant choosing a sooner loss. Most of the predicted probabilities for later gains are below 0.5 although there are some points above the 0.5 line.

Table 5.16: Table of multilevel logistic regression results.

	Regression coefficients: Log odds (95% confidence interval)				
Terms	Null model	Model 1	Model 2	Model 3	Model 4
Intercept	-0.65 (-0.81, -0.50)	-0.19 (-0.38, 0.01)	0.13 (-0.55, 0.80)	0.30 (-0.43, 1.03)	0.43 (-0.32, 1.17)
(sign)loss		-1.02 (-1.27, -0.78)	-1.02 (-1.27, -0.78)	-1.02 (-1.27, -0.78)	-1.30 (-1.72, -0.89)
amount ratio			-0.39 (-1.19, 0.41)	-0.39 (-1.19, 0.41)	-0.39 (-1.19, 0.41)
(gender)M				-0.21 (-0.54, 0.12)	-0.21 (-0.54, 0.12)
(education)Bachelors and above				0.36 (0.05, 0.68)	0.36 (0.05, 0.68)
(marital)Cohabit/Sep				-0.10 (-0.58, 0.39)	-0.10 (-0.69, 0.49)
(marital)Married				-0.49 (-0.85, -0.14)	-0.73 (-1.16, -0.31)
(sign)loss:(marital)Cohabit/Sep					0.01 (-0.75, 0.77)
(sign)loss:(marital)Married					0.56 (0.02, 1.10)
Random effects					
τ_{00}	0.15	0.19	0.19	0.12	0.12
ICC	0.04	0.05	0.05	0.04	0.04
Participants	64	64	64	64	64
Observations	1,280	1,280	1,280	1,280	1,280
AIC	1,648.8	1,580.9	1,582.0	1,578.3	1,577.4

Reference categories for sign is gain, gender is female, for education is below Bachelors, and for marital status is single.

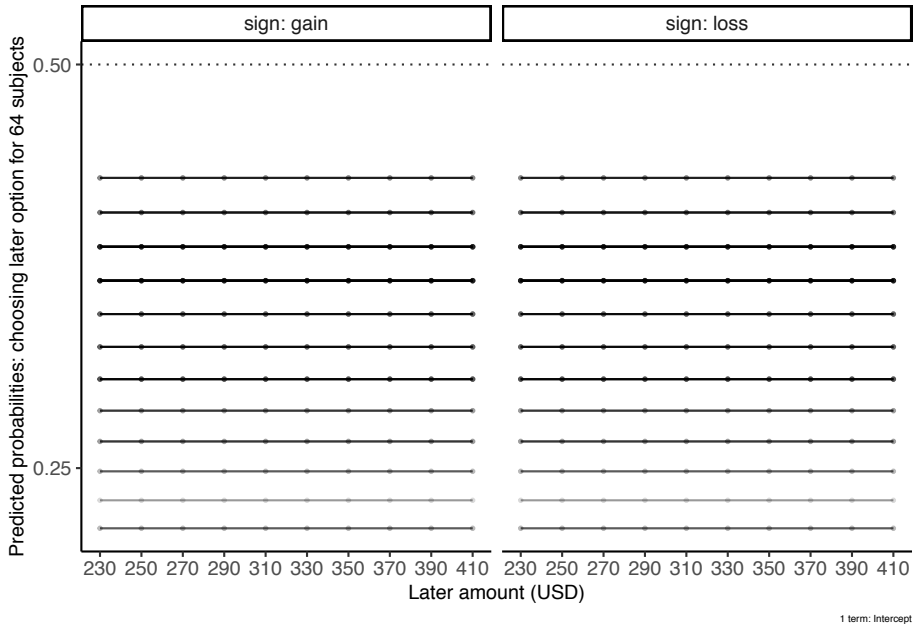


Figure 5.63: Predicted probabilities for null model.

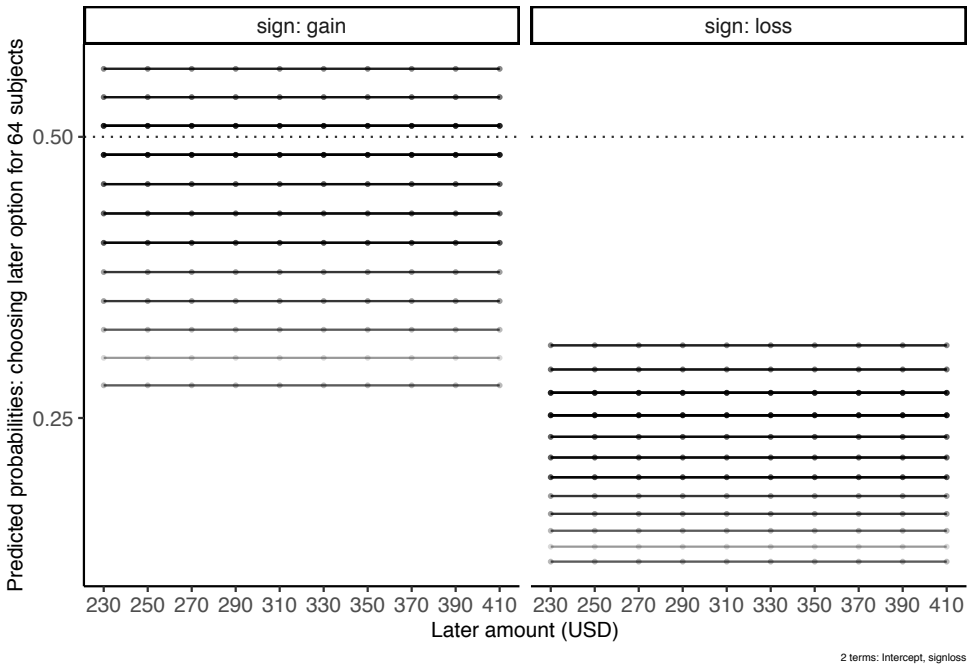


Figure 5.64: Predicted probabilities for model 1.

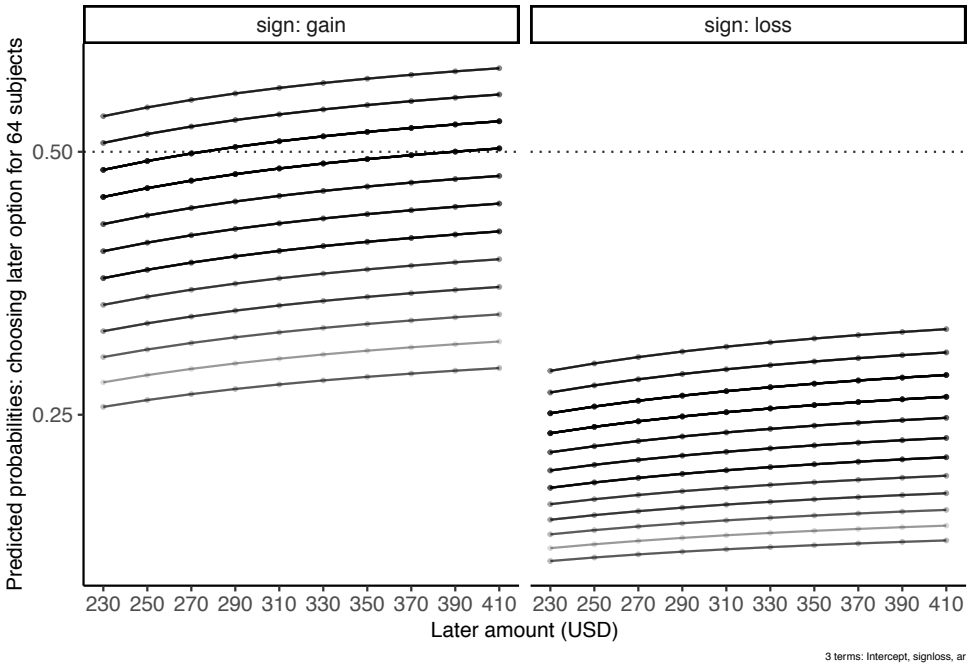


Figure 5.65: Predicted probabilities for model 2.

5.5.5 Diagnostics

Table 5.17 presents the diagnostic results for model 2. It provides an additional assessment of the model. The results are separated for gain questions only, loss questions only, and for all questions (both gain and loss).

Table 5.17: Diagnostics results for model 2. The later choice is taken as a "positive", while the sooner choice, a "negative".

Sign	Later choices	Sooner choices	Sensitivity (%)	Specificity (%)	Prevalence (%)	PPV (%)	NPV (%)
all	445	835	37.3	95.9	34.8	83	74.2
gain	292	348	56.8	90.2	45.6	83	71.4
loss	153	487	0.0	100.0	23.9	NaN	76.1

Overall, model 2 is able to correctly predict the later choice less than 40% of the time (sensitivity) and the sooner choice almost all of the time (specificity). The model is able to correctly predict later gains 57% of the time and the sooner gain 90% of the

time. For losses, it is able to correctly predict the sooner loss every time. However, it is never able to correctly predict the later loss. This can be seen from the predicted probabilities for losses, which are all below 0.5.

For the given prevalence, overall, the model is able to correctly predict the sooner (NPV) and later (PPV) choices about 74% and 83% of the time respectively. The model is able to correctly predict the sooner and later gains about 71% and 83% of the time respectively. The model is able to correctly predict the sooner loss 76% of the time. It is not able to calculate a PPV for losses as the sensitivity is zero and the specificty is 100%, which makes the denominator zero during the calculations.

5.6 Hardisty and Weber (2009) Experiment 2

This section focusses on Experiment 2 of Hardisty and Weber (2009) and follows from the previous section, which covered Experiment 1. There are 116 participants in Experiment 2 (56% female). Each participant answered 10 questions on gaining money and 10 questions on losing money. Questions for gains and losses were similar in all respects except for their sign, i.e. whether the monetary amounts were framed as a gain or a loss. The sooner amount was always kept constant at USD250, while the later amount varied from USD230 to USD410, increasing by USD20 each time.

5.6.1 Exploratory data analysis

Although there are other demographic characteristics provided by the original study author in an Excel file, no data dictionary or description is provided for the variables and how they were coded. Thus, only age and gender will be used here.

The median age of the participants was 39 years old (range: 18 to 77 years). The median age for men was 38 years old. The median age for women was 39 years old.

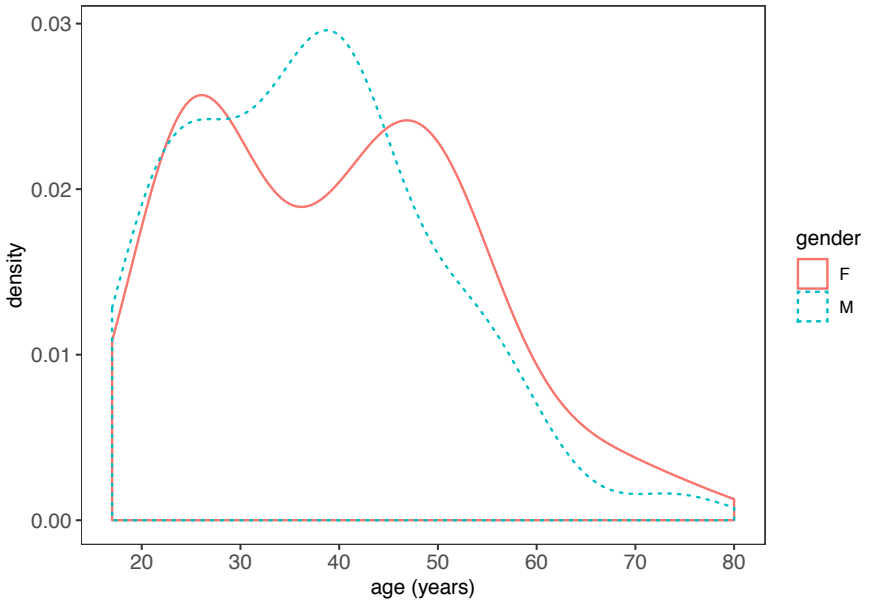


Figure 5.66: Density plots of age for men and women.

Figure 5.66 shows the distribution of age by gender. The mode for men is around 40 years old. The distribution for women appears bi-modal with a peak between 20–30 years and another peak between 40–50 years.

Figure 5.67 shows the proportion of later choices for each participant. Each point represents the overall proportion of later choices for one participant. The colour and shape of the points indicate whether the proportion is for gains (blue circle) or losses (red crosses). There is a dotted vertical line at 0.5 on the x-axis to highlight the points to the left and right of the line.

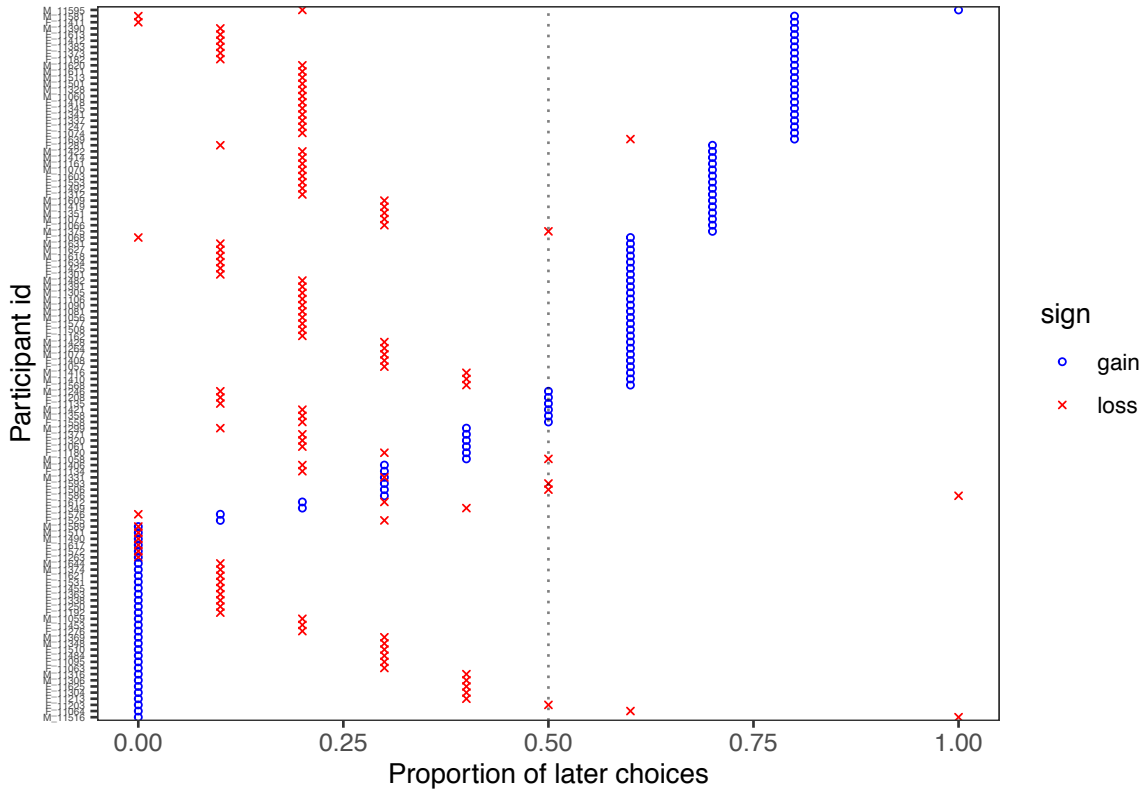


Figure 5.67: Strip chart showing the proportion of later choices for each participant. The participants on the x-axis are ordered by increasing proportion of later gains. The colour and shape of the points represent the two different values of sign: gain and loss.

From Figure 5.67, two-thirds (67%) of the proportions are smaller than 0.5, which implies that most choices are for the sooner option. For losses, almost all (92%) of the proportions are smaller than 0.5. For gains, just under half (41%) the proportions

are smaller than 0.5. Two-thirds (66%) of participants choose the later gain more often than the later loss, which can be seen from the red crosses to the left of blue circles for most participants.

Figure 5.68 shows the distribution of the proportion of later choices by gender and sign. The distribution for gains are on the left panel and losses on the right panel. There were 232 observations in total. Each participant had two observations, one for the proportion of later choices for gains and the other for losses.

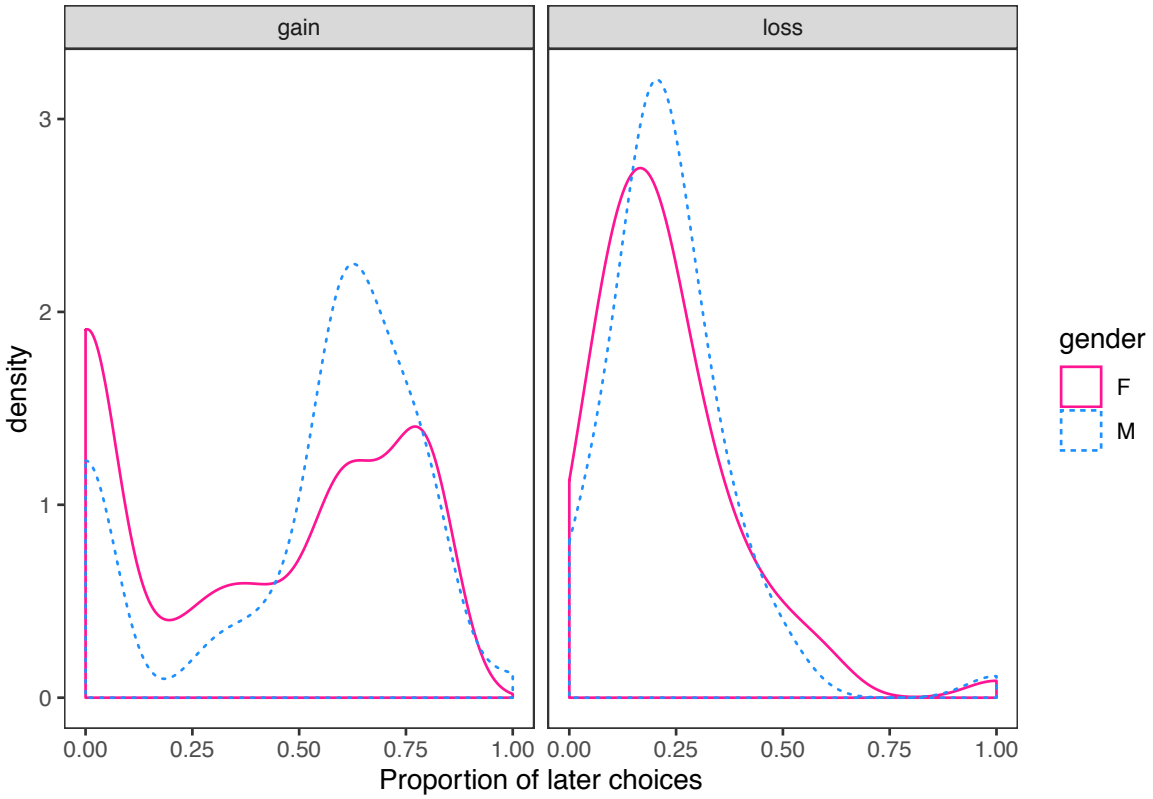


Figure 5.68: Density plots of the proportion of later choices by sign, coloured by gender.

The overall shape of the density plots are similar for men and women. For gains, the densities were bimodal for both men and women. However, women had more values between 0 and 0.2 on the x-axis, indicating that men tended to take the later gain slightly more often than women. For losses, both densities had a tall peak between 0 and 0.5.

Figure 5.69 shows the relationship between the proportion of later choices and age by sign. Each point is one observation, with darker points representing overlapping observations. A blue loess line is drawn to display any trend. There does not appear to have any relationship between the proportion of later choices and age.

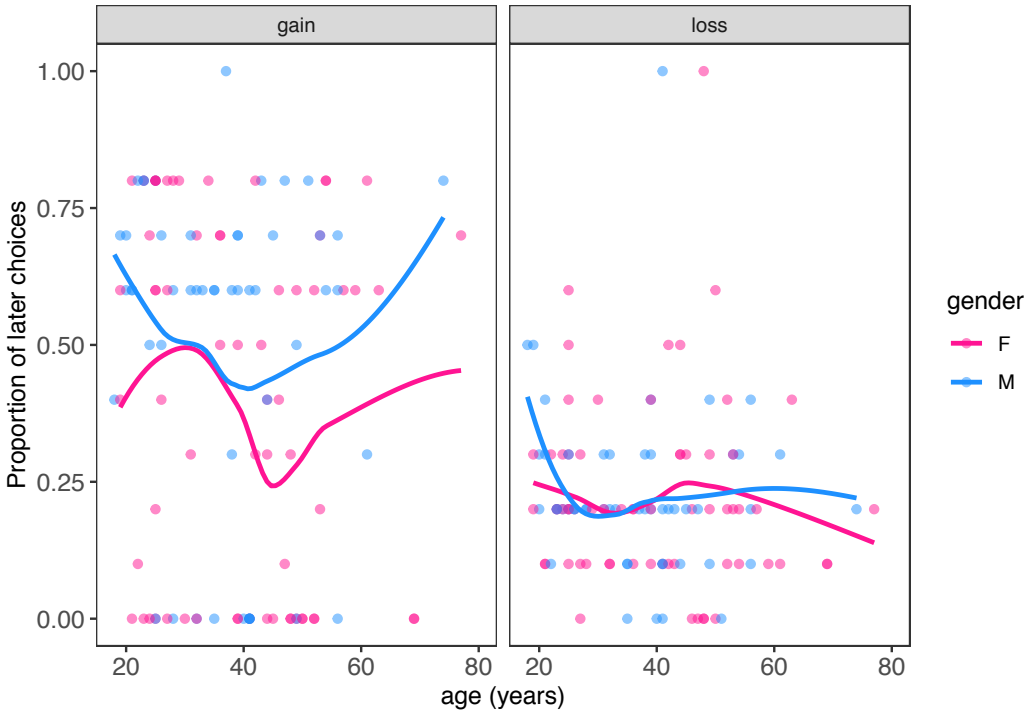


Figure 5.69: Scatter plots of the proportion of later choices and age by sign with a smooth blue trend line.

Figure 5.70 shows the relationship between the proportion of later choices and amount ratio for men and women. Both men and women have similar patterns for gains and losses. For gains, as the amount ratio increases, the proportion choosing the later gain decreases and reaches zero when the amount ratio is 1. For losses, the proportion of choosing the later loss increases as the amount ratio increases. However, unlike in Study 1, the proportion of later losses does not reach 1 when the amount ratio is 1 or even when the amount ratio is close to 1.1. There is still a number of participants who choose the sooner loss when the sooner and later amounts are the same and even when the sooner amount is larger in magnitude than the later amount.

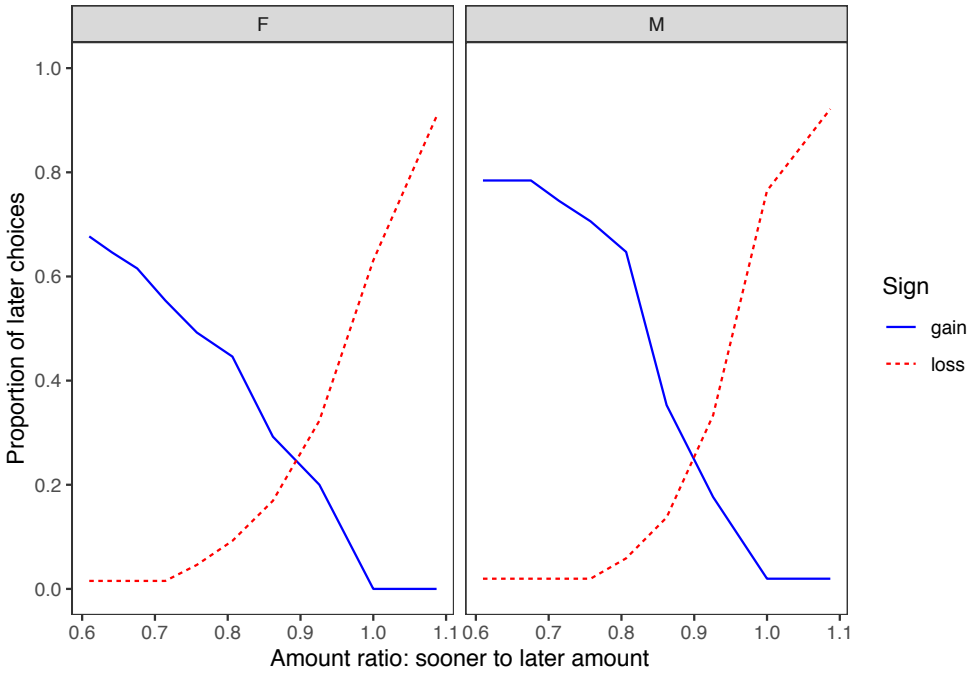


Figure 5.70: Line plots of the relationship between the proportion of later choices and amount ratio by gender. Lines are coloured by sign.

5.6.1.1 Indifference points

Participants had either zero or one switching point because 37 (22%) participants were excluded on the basis that they 'failed' the study's 'careful-response criteria' (Hardisty and Weber 2009, 334).

Figure 5.71 shows each choice each participant made for the 10 different amounts of money available later. A black circle represents a sooner choice and a green triangle a later choice.

There were 33 (28%) participants who were completely consistent in their choices for gains (men: 24%; women: 32%), i.e. they always chose the sooner or later gains. For losses, 12 (10%) participants were completely consistent in their choices (men: 10%; women: 11%). There were 7 participants who were completely consistent for both gains and losses.

Figure 5.72 shows the densities of the indifference points of participants coloured

by sign. The indifference point is calculated by adding the later amount where participants first switched preferences (e.g. from preferring the later to the sooner option or vice versa) and the later amount before the first switch and then dividing this by 2. Participants who did not switch and always chose the sooner amount or the later amount were given the minimum and maximum values of the later amount, i.e. 230 and 410, respectively.

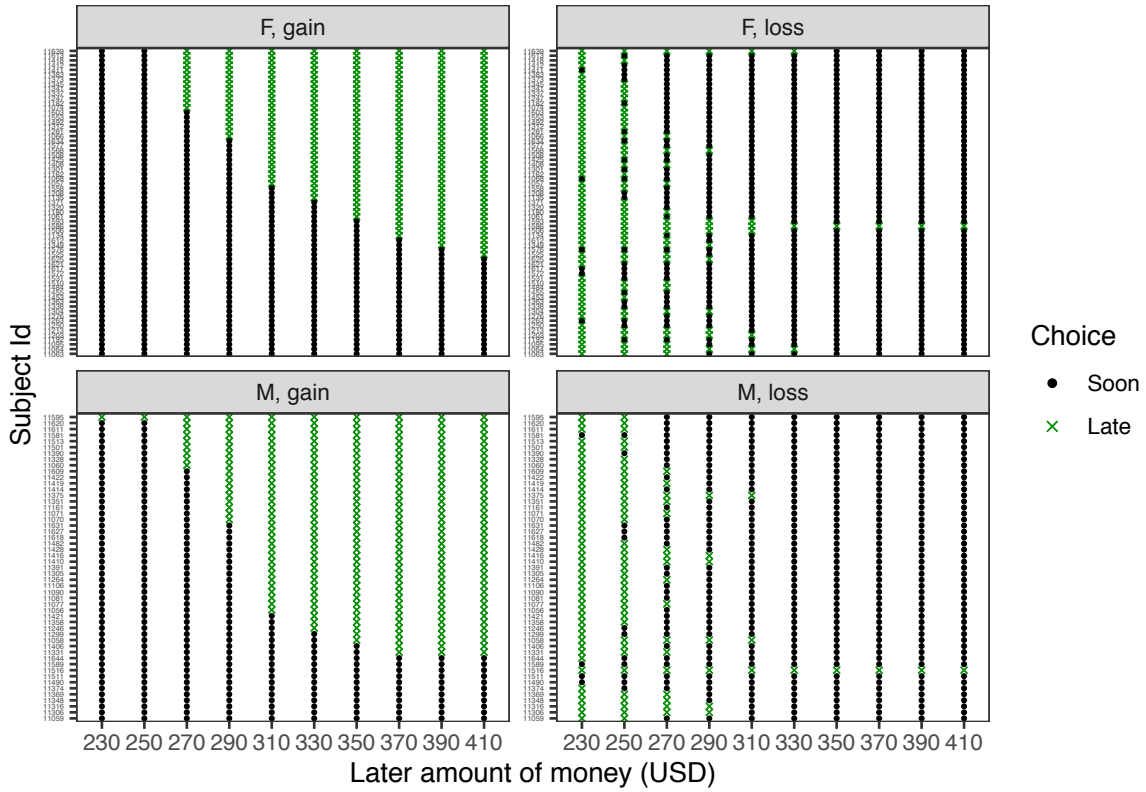


Figure 5.71: Strip charts of each choice each participant made by sign. Participants are ordered by decreasing number of later choices for gains.

For losses, the peaks tend to occur at the smallest values (e.g. 240, 260, 280), with no peaks from 360 to 400. For gains, there is no peak at 240 but there are tall peaks from 260 to 300 and smaller peaks after. This suggests that participants tend to switch preferences earlier for losses than gains. The taller peak for gains at 230 suggests a higher proportion of participants who always chose sooner gains than sooner losses.

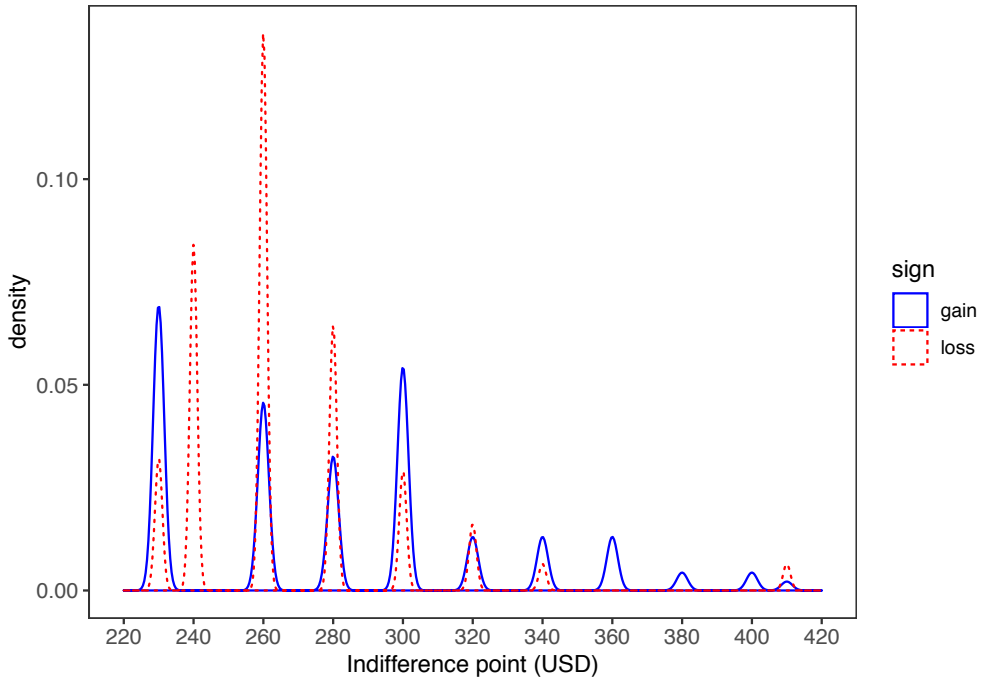


Figure 5.72: Density plots of the indifference points coloured by sign.

There were 38 participants (33% of all 116 participants) who always chose the later option and thus did not have a switching point, of which 7 participants did not switch for both gains and losses. These 38 participants are reflected by the tall peak at 230 and small peak at 410.

5.6.2 Statistical modelling

There is evidence that the between-subject variance is non-zero. The likelihood ratio statistic for testing the null hypothesis that the variance of the average subject is zero, i.e. $\sigma^2_{u0} = 0$, can be calculated by comparing the two-level null model with the correspondong single-level model. The test statistic is 83.9 with 1 degree of freedom from a chi-squared distribution.

Figure 5.73 shows the estimated residuals for all 116 participants in the sample. A few participants have a 95% confidence interval that does not overlap the horizontal line at zero. For these few participants, the choices for the later option are significantly

above or below average (above/below the zero line).

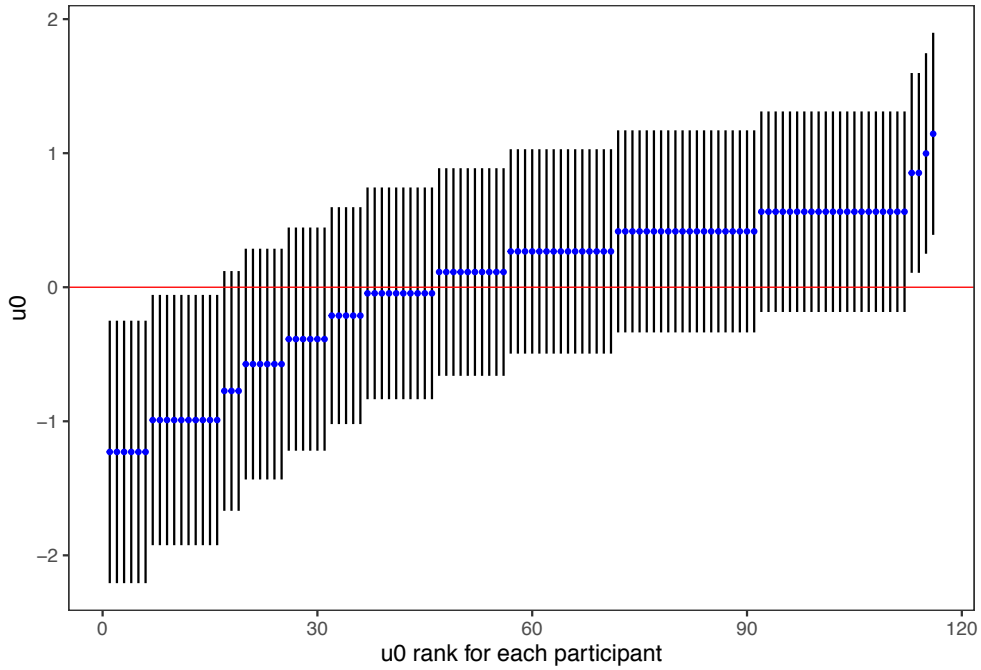


Figure 5.73: Estimated random effects residuals for each participant.

Results from 6 models are displayed in Table 5.18. The models were selected based on the findings from the exploratory data analysis. The estimated coefficients are rounded to two decimal places for presentation purposes. Sign and gender are categorical variables. The reference category for sign is gain and for gender is women.

Across the models, the intraclass correlation coefficient (ICC) ranges between 0.14 and 0.16. This means about 14–16% of the variance is explained by the grouping structure in the population. In Table 5.18, τ_{00} represents the between-subject variance.

None of the models performed particularly well based on the AIC values. The model with only the intercept and sign performed well relative to the other models with an AIC of 2,743. Adding in other terms, including the amount ratio, only improved the model slightly. Models with an interaction between amount ratio and sign produced unreliable results, i.e. very large coefficients.

Table 5.18: Table of multilevel logistic regression results.

	Regression coefficients: Log odds (95% confidence interval)					
Terms	Null model	Model 1	Model 2	Model 3	Model 4	Model 5
Intercept	-0.78 (-0.94, -0.62)	-0.29 (-0.48, -0.10)	-0.51 (-1.04, 0.03)	-0.64 (-1.20, -0.09)	-0.90 (-1.62, -0.18)	-0.73 (-1.29, -0.17)
(sign)loss		-1.09 (-1.29, -0.90)	-1.10 (-1.29, -0.90)	-1.10 (-1.29, -0.90)	-1.10 (-1.29, -0.91)	-0.90 (-1.15, -0.64)
amount ratio			0.27 (-0.35, 0.88)	0.27 (-0.35, 0.88)	0.58 (-0.25, 1.42)	0.27 (-0.35, 0.88)
(gender)M				0.32 (-0.02, 0.66)	0.88 (-0.18, 1.93)	0.50 (0.12, 0.87)
(gender)M:amount ratio					-0.69 (-1.92, 0.55)	
(sign)loss:(gender)M						-0.43 (-0.81, -0.05)
Random effects						
τ_{00}	0.52	0.61	0.61	0.58	0.59	0.58
ICC	0.14	0.16	0.16	0.15	0.15	0.15
Participants	116	116	116	116	116	116
Observations	2,320	2,320	2,320	2,320	2,320	2,320
AIC	2,874.4	2,743.2	2,744.5	2,743.1	2,743.9	2,740.3

Reference categories for gender is female and for sign is gain.

5.6.3 Predicted probabilities

The predicted probabilities are constant when the amount ratio term is not included. The predicted probabilities for later losses are almost always below 0.5, which implies that the models almost always predict a participant choosing a sooner loss. Most of the predicted probabilities for later gains are below 0.5 although there are some points above the 0.5 line.

Although not shown, in models 3 and 4, the intercepts for women tend to be decreasing over the later amounts while the intercepts for men tend to increase very slightly. However, across the models, the predicted probabilities appear similiar as almost straight lines.

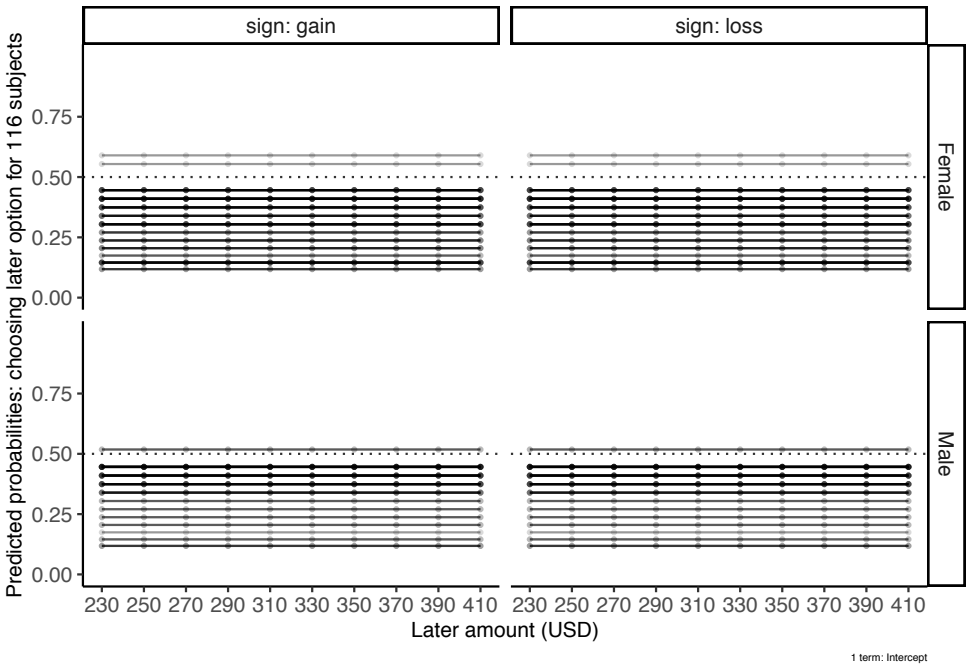


Figure 5.74: Predicted probabilities for null model.

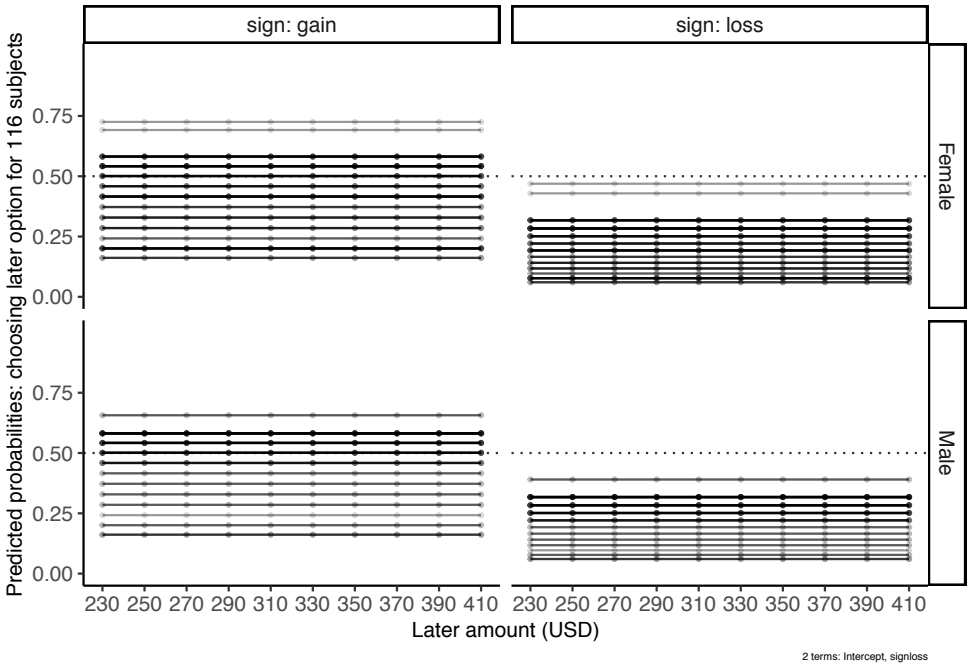


Figure 5.75: Predicted probabilities for model 1.

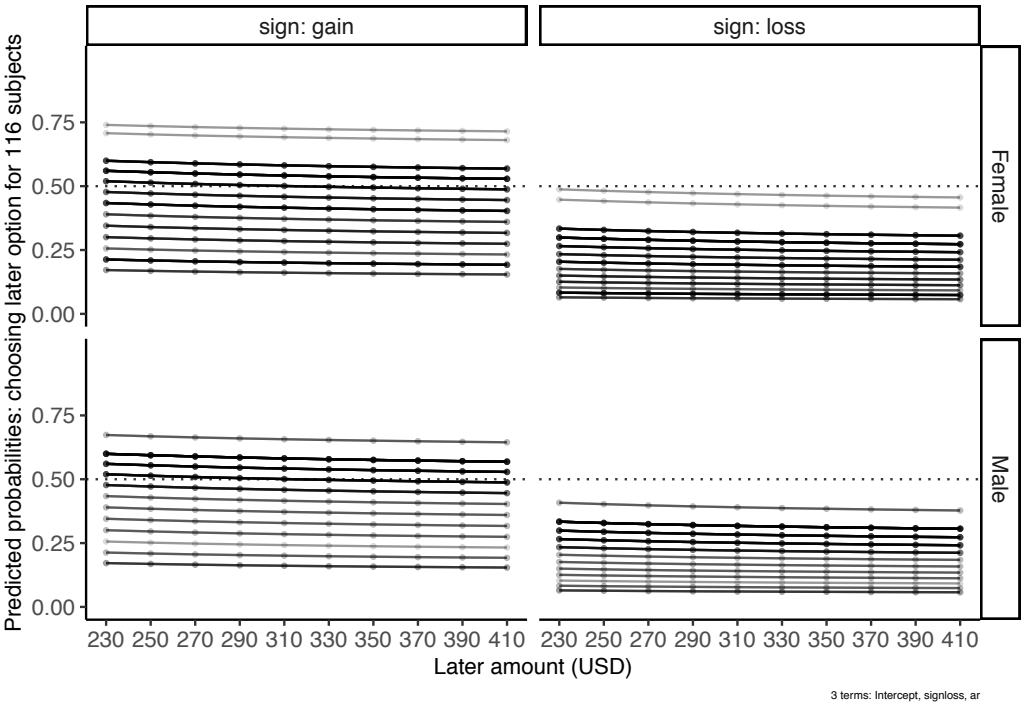


Figure 5.76: Predicted probabilities for model 2.

5.6.4 Diagnostics

Table 5.19 below presents the diagnostic results for model 2. It provides an additional assessment of the model. The results are separated for gain questions only, loss questions only, and for all questions (both gain and loss).

Table 5.19: Diagnostics results for model 2. The later choice is taken as a "positive", while the sooner choice, a "negative".

Sign	Later choices	Sooner choices	Sensitivity (%)	Specificity (%)	Prevalence (%)	PPV (%)	NPV (%)
all	774	1,546	43.0	87.6	33.4	63.4	75.4
gain	511	649	65.2	70.4	44.1	63.4	72.0
loss	263	897	0.0	100.0	22.7	NaN	77.3

Overall, model 2 is able to correctly predict the later choice less than half of the time (sensitivity) and the sooner choice close to 90% of the time. The model is able to correctly predict sooner and later gains about two-thirds of the time. For losses, it is able to correctly predict the sooner loss every time. However, it is never able to correctly predict the later loss. The model does not have any true positives or false positives.

For the given prevalence, overall, the model is able to correctly predict the sooner (NPV) and later (PPV) choices about 63% and 75% of the time respectively. The model is able to correctly predict the sooner and later gains about 72% and 63% of the time respectively. The model is able to correctly predict the sooner loss 77% of the time. It is not able to calculate a PPV for losses as the sensitivity is zero and the specificty is 100%, which makes the denominator zero during the calculations.

5.7 Discussion

This section synthesises results from the chapter. It plots various results obtained from the multilevel models to examine consistencies and patterns across studies (Figures 5.77 and 5.78). Then, it summarises findings uncovered in the chapter such as, the level of heterogeneity within and between studies, as well as the difficulties in estimating an indifference point for each participant arising from zero or multiple switching points. Contributions and limitations are discussed.

The aim of this chapter is to understand the factors that influence individual participants' choices in the presence of gains and losses. Results from the multilevel models provide evidence of a 'sign' effect, where individual choices were influenced by whether the amounts offered were presented as gains or losses. In five of the six studies analysed, the odds of choosing the later amount decreased when it was presented as a loss, compared to a gain. In five studies, there was a negative association between the amount ratio and choosing the later option, which means that as the amount ratio increases the odds of choosing the later option decreases.

Figure 5.77 shows the odds ratio with confidence interval of the sign term from two different models in each study. These models were chosen for comparison as the terms were common across all studies. One model contains only sign and the intercept while the other model has amount ratio in addition to sign and intercept. The size of the points is proportional to the number of participants, with larger points indicating more participants.

An odds ratio (OR) of 1 indicates no effect of the sign term. An OR smaller than 1 indicates lower odds of choosing the later loss compared to gains. An OR bigger than 1 indicates higher odds of choosing the later loss compared to gains. The estimates from the two models overlap.

The multilevel models did not perform particularly well in terms of predicting choices. Figure 5.78 shows the sensitivity and specificity of the "best" model for each study. The multilevel models performed poorly at predicting later losses as the sensitivities for losses were mostly close to zero. It was able to predict the sooner loss (specificity)

very well partly because most choices for losses were for the sooner amount.

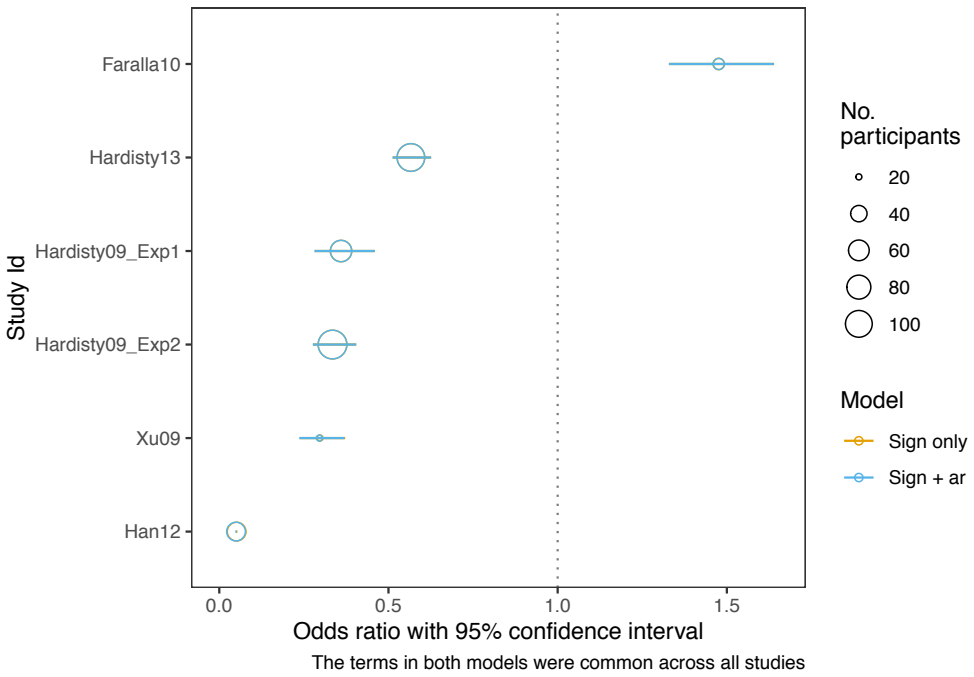


Figure 5.77: Odds ratio with confidence interval of the sign term from two models in each study that had common terms. One model contains only the sign and intercept while the other model includes the amount ratio. The size of the points is proportional to the number of participants. The papers are ordered by increasing odds ratio.

Based on the results in previous sections of this chapter, the multilevel models also provide evidence of substantial heterogeneity within and between studies. For example, the between-subject variance ranged from 0.01 to 0.61 across studies, while the intraclass correlation ranged from 0.01 to 0.16 across studies. Although the data were analysed with single-level models in the original studies, there was consistent evidence that the between-subject variance was non-zero, which suggests that it is more appropriate to analyse the data using multilevel models.

Even meaningful associations at the aggregate level tended to be inconsistent at the individual level. There was substantial heterogeneity between participants within a study. It was common to find a substantial number of participants displaying no or even the opposite association to that observed at the aggregate level. There was also

substantial heterogeneity in the associations observed at the individual level across the different studies.

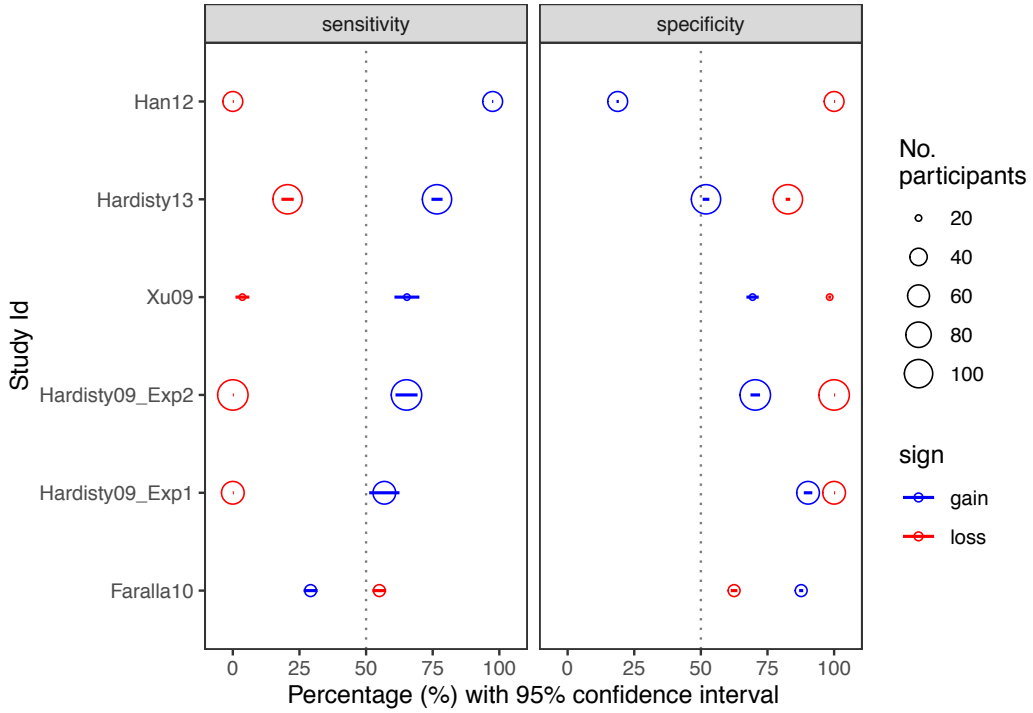


Figure 5.78: Sensitivity and specificity with confidence interval from the "best" model for each paper. Papers are ordered by increasing percentage for sensitivity of gains. Size of points is proportional to the number of participants.

It is difficult to compare any other terms in the models across studies as the outcome sign and amount ratio were the only two common terms. Studies had factorial designs but manipulated different factors and the associated levels. For example, some studies manipulated the delay of the later amount, others manipulated the delays of the sooner and later amount while others did not manipulate delay at all. Within a study, delays could all be relatively short or range from 1 week to 25 years. However, there was no consistent association between delay and choosing the later amount.

Although the studies had factorial designs, which could be used to estimate indifference points for the participants, not all reported analysing indifference points (Xu et al. 2009; Faralla et al. 2012). In both studies, only a third of all possible indifference

points could be calculated as a result of participants switching more than once or not switching at all in their choices. The number of indifference points that could be calculated was higher in the other studies partly because up to a third of participants who 'switched back and forth more than once' or were 'inattentive' during the study were dropped and not included in the dataset provided (Hardisty and Weber 2009, 331). However, there were still about a third of the remaining participants in the study who did not switch preferences at all.

It was surprising to uncover the large number of uncalculable indifference points for each participant as such information is not typically reported in the literature. There were studies where many participants had 50% of indifference points that could not be calculated. Given the prevalence of uncalculable indifference points, it is worth asking if analyses using discount rates are still appropriate. Without a unique indifference point, a discount rate cannot be estimated. It is also worth highlighting that it is common for studies to drop participants who switch more than once or to impute a value for for participants who did not switch at all (Hardisty et al. 2013). For example, depending on the design, the minimum value of the sooner or later amount could be used for participants who always chose the sooner option and the maximum value for participants who always chose the later option. However, this inappropriately assumes that everyone *should* have a single switching point.

Given the substantial heterogeneity in the data, it would not be advisable to conduct a one-stage meta-analysis based on aggregate data or results reported in studies. Such an approach is common in behavioural science, e.g. a recent meta-analysis of intertemporal choice used aggregated estimates reported in studies (Imai, Rutter, and Camerer 2019). However, results from this chapter suggest that aggregated estimates should be treated with caution as they can be inaccurate and they do not account for the substantial heterogeneity within and between participants.

5.7.1 Contributions

This chapter provides the first in-depth analysis of choices on question pairs involving gains and losses using individual participant data (IPD) from multiple studies.

Across the six studies with IPD, there was no consistent association between individual choices and whether the amounts were presented as gains or losses. There was no common variable that had a consistent association with individual choices across the studies, partly due to the different study designs.

Results provide evidence that individuals are substantially more heterogeneous, and there is more between-subject variance within a study, than discussed in the literature. This chapter demonstrated the use of multilevel models, which are more appropriate than the single-level models commonly used in the literature. The within-subject heterogeneity and the between-subject variance in the studies call into question results reported in the literature, which are mostly based on single, aggregate-level analyses.

The plots used in this chapter to visualise associations at the individual and question level provide a graphical template for researchers in the field. These plots visualise within and between subject heterogeneity, which can inform subsequent statistical modelling choices. This includes visualising the predicted probabilities of each participant from the multilevel model when the study have complex factorial designs.

Finally, our findings on the sign effect has broader implications for the reproducibility debate. The Open Science Collaboration (2015) and Camerer et al. (2016) attempted to replicate the original studies chosen by following the methods and analysis procedures as closely as possible. Our findings call into question the validity of this method of replication and suggest that the current 'reproducibility crisis' may stem from more foundational issues that need addressing. Replication attempts should be more critical, they should critique definitions and explore assumptions and heterogeneity to inform any subsequent analysis.

5.7.2 Limitations

Results from this chapter may not generalise to the intended population of all studies that had question pairs for gains and losses as individual participant data from only six studies were analysed. The six studies were from a convenience sample of authors

who responded to emails requesting data for a more general purpose of creating a public database of binary choice questions. Three studies were also from the same author (Hardisty and Weber 2009; Hardisty et al. 2013). There was also insufficient information to determine if the original participants were drawn from a random sample. If they were from a convenience sample, then it would be unclear to what population these results generalise.

The multilevel models could be improved by including a random slope term. However, the current models are already limited by the small sample of participants. While there are many more questions per participant, the unit of analysis is at the individual level. This means that the effective sample size is the number of participants and not number of responses to questions. Given the limited sample size, models could be simplified or made more complex only if there were more participants.

A checklist could be used to further appraise the quality of evidence. This will provide a broader overview of the strengths and weaknesses of the studies. However, there is still ongoing discussions as to what would be the most appropriate checklist to complement IPD analyses in the behavioural sciences. If this chapter were to progress into a full systematic review, then a suitable checklist should be used.

Finally, results from the multilevel models are based on exploratory, and not confirmatory, analyses. The terms in the models were not pre-specified and multiple combinations of the terms were fitted. As such, the estimated coefficients and confidence intervals should be treated with some caution. Future work can build on the results by designing and analysing confirmatory studies to test the associations identified in this chapter.

Chapter 6

Conclusion

The initial aim of this thesis was to conduct a systematic review and meta-analysis of intertemporal choice anomalies. In medicine, the gold standard of conducting a quantitative review involves appraising the quality of included studies with a standardised checklist. However, this thesis did not find suitable checklists in behavioural science.

In Chapter 2, a checklist of recommended statistical practices was developed. The items in the checklist were mainly informed by guidelines that were created by a team of behavioural scientists and statisticians for the American Psychological Association (Wilkinson and Task Force on Statistical Inference 1999). The checklist items were also, albeit to a lesser extent, informed by guidelines in neighbouring disciplines such as animal studies, medicine and neuroscience. Then, a team of statisticians and behavioural scientists used the checklist to evaluate a sample of economics and psychology studies that have been independently replicated.

Results from Chapter 2 contribute to the debate surrounding the 'reproducibility crisis' in behavioural science. On average, the evaluated studies adhered to 30% of recommended statistical practice. Incomplete reporting hampered meaningful evaluation of the association between the checklist scores and replication success. Ethical implications and reproducibility concepts were discussed. Improving the quality of reporting in behavioural science will facilitate replication efforts and the

checklist in this chapter provides a standardised template of essential information that should be reported.

Chapters 3 and 4 focusses on the sign effect, which is an established intertemporal choice anomaly. The sign effect refers to the observation that people tend to discount gains more than losses. In chapter 3, verbal descriptions of the sign effects were critically discussed and shown to be ambiguous. The concept of a discount rate, which is used to analyse the sign effect, was also critically discussed. The descriptions of the sign effect were formalised and a hypothesis testing framework was proposed.

Chapter 4 details the attempt to conduct the first systematic review and meta-analysis of the sign effect anomaly, having defined it more mathematically. The data set had a mix of question-level and individual participant data from nine studies. Results provided evidence for a sign effect at the aggregate, question-level and individual-level. However, analysing the sign effect at the individual-level revealed substantial heterogeneity not discussed in the literature, including the issue that only about 30% of observations can be used to estimate the sign effect in study. Most observations involved no discounting of losses or gains. Given the lack of informative responses, formal statistical modelling of the sign effect was not undertaken.

Chapter 5 detailed the formal statistical modelling of individual participants' responses to questions involving gains and losses. The models estimated the extent to which, along with other factors, the outcome sign, i.e. whether amounts were presented as gains or losses, influenced choices. This contributes to the goal of understanding how intertemporal choices are made (Read 2004).

Results from the multilevel models suggest that the later amount was chosen more often for gains than losses. However, the models did not perform particularly well in accurately predicting choices, especially later losses as participants tended to almost always prefer the sooner loss. There was also substantial heterogeneity between and within participants. There was a surprisingly large number of indifference points that could not be calculated.

In light of the 'reproducibility crisis' in behavioural science, researchers are attempting to replicate more published studies and encourage prospective studies to be

pre-registered. However, this thesis demonstrated several other key issues that need to be considered. For example, standardised reporting of essential information will facilitate replication efforts. This involves having accepted guidelines that authors and reviewers adhere to, which requires a concerted long-term effort (Altman and Simera 2016). Definitions need to be explicitly formalised, data need to be described sufficiently, assumptions need to be explored empirically, study designs need to be informative and the different types of heterogeneity need to be accounted for. Good statistical practice is central to tackling the 'reproducibility crisis' in behavioural science. The opportunity is ripe for statisticians and behavioural scientists to collaborate in a similar vein to Wilkinson and Task Force on Statistical Inference (1999).

Chapter 7

Appendix

7.1 Search terms

1. Search engine: Scopus

Searched: SUBJAREA (ECON) OR SUBJAREA (PSYC) AND (TITLE(guideline*) OR TITLE(checklist*) OR TITLE(recommend*)) AND (TITLE(statistic*) OR TITLE(method*) OR
TITLE(reporting))

Results: 194 records found

Searched: (SRCTITLE (*psycholog*) OR SRCTITLE (*behavior*) OR SRCTITLE(*brain*) OR SRCTITLE (*cogniti*) OR SRCTITLE (*personality*) OR SRCTITLE(*emoti*) OR SRCTITLE (*econom*)) AND (TITLE (guideline*) OR TITLE (checklist*) OR TITLE(recommend*)) AND (TITLE(statistic*) OR TITLE(method*) OR TITLE(reporting))

Results: 153 records found

2. Search engine: Web of Science

Searched: SU=(Behavioral Sciences OR Business & Economics OR Psychology) AND TI=(guideline* OR checklist* OR recommend*) AND TI=(method* OR statistic* OR reporting)

Results: 443 records found

3. Search engine: PSYCINFO [Limited to certain fields of psychology]

Searched: (ti(guideline* OR checklist* OR recommend*)) AND (ti(statistic* OR method* OR reporting)) AND cl("Human Experimental Psychology" OR "Statistics & Mathematics" OR "Research Methods & Experimental Design" OR "Organizational Behavior" OR "Consumer Psychology" OR "Educational Psychology" OR "Social Psychology" OR "General Psychology" OR "Psychometrics & Statistics & Methodology" OR "Psychosocial & Personality Development" OR "Visual Perception" OR "Industrial & Organizational Psychology" OR "Attention" OR "Health Psychology & Medicine" OR "Educational Psychology")

Results: 78 records found

4. Search engine: Econlit with full text

Searched (boolean/phrase): TI (guideline* OR checklist* OR recommend*) AND TI(method* OR statistic* OR reporting)

Results: 61 records found

5. Search engine: ScienceDirect

Searched: TITLE(guideline* OR checklist* OR recommend*) AND TITLE(method* OR statistic* OR reporting)[All Sources(Economics, Econometrics and Finance, Psychology)]

Results: 144 records found

7.2 The 55 checklist items used to calculate the score

1. "Were hypotheses, aims, or goals clearly stated?"
2. "If study had multiple goals, were they defined and prioritised?"
3. "Use your judgment: was sample representative of target population?"
4. "Did authors explicitly define each outcome variable?"
5. "Did authors explicitly describe how each outcome variable relates to goals of study?"
6. "Did authors explicitly describe how each outcome variable was measured?"
7. "If an instrument was used to collect data, did author(s) describe the reliability with regard to the way the instrument is used in a population?"
8. "If an instrument was used to collect data, did author(s) describe the validity with regard to the way the instrument is used in a population?"
9. "If a physical apparatus was used, did author(s) describe the brand?"
10. "If a physical apparatus was used, did author(s) describe the model?"
11. "If a physical apparatus was used, did author(s) describe the design specifications?"
12. "Did authors report planning sample size in advance?"
13. "Was sample size reported?"
14. "Did author(s) describe any anticipated sources of attrition due to noncompliance, dropout, death, or other factors?"
15. "Did author(s) describe how such attrition may affect the generalisability of results?"

16. “Were pilot study sessions conducted?”
17. “If pilot sessions were conducted, was the purpose stated?”
18. “Did author(s) describe personnel who collected the data?”
19. “Did author(s) describe personnel who administered the study?”
20. “Did author(s) explicitly specify randomising treatment to participants?”
21. “Use your judgment: was allocation adequately concealed?”
22. “Were participants informed of the true purpose of the study?”
23. “If deception was used, was it justified by authors?”
24. “Was there monetary incentive?”
25. “Were responses automatically captured, e.g. computer software programme?”
26. “If responses were not automatically captured, was an attempt made to blind personnel collecting data?”
27. “Was an attempt made to blind participants to the treatment group assigned?”
28. “Was an attempt made to blind personnel administering the study?”
29. “Was an attempt made to blind those analysing the data?”
30. “Were methods used to handle multiple testing?”
31. “Did author(s) describe participants excluded/dropped/lost to follow up?”
32. “Were analyses compared with and without the participants excluded/dropped/lost to follow up?”
33. “Did author(s) describe the differences with and without the participants excluded/dropped/lost to follow up?”
34. Did author(s) report the demography / characteristics of each group?

35. “Were groups tested for baseline differences?”

36. “Was the number of of participants in each group stated?”

37. “Was sample size equally balanced across groups (+/- 10% difference)?”

38. “Use your jugment: was there evidence that data quality was checked (e.g. outliers, illegal values, anomalies in the data)?”

39. “Use your judgement: was basic data adequately described? (e.g. What was the distribution tf the data? Was M +/- SD provided for normally distributed data? Was interquartile range or graphics provided of skewed data?)”

40. “Was each statistical method described sufficiently to understand what was done? E.g. Spearman’s rank correlation and not simply ‘correlation’ ”

41. “Did author(s) justify the use of each statistical method?”

42. “Did author(s) describe the underlying assumptions of each analysis?”

43. “Use your judgment: do underlying assumptions seem reasonable given the data?”

44. “Were residuals examined?”

45. “Were residuals presented graphically?”

46. “Was a protocol/pre-registered study mentioned?”

47. “Was there any deviation from the protocol/pre-registered study?”

48. “Use your judgment: were all outcomes tested, reported?”

49. “Were actual p-values reported (e.g. 0.035 instead of $p<0.05$) for each finding except where the p-value is less than 0.001?”

50. “Were effect sizes presented for each finding?”

51. “Were interval estimates presented for each finding?”

52. “Were interval estimates presented in each figure, where appropriate?”
53. “Did the author(s) describe the statistical software used?”
54. “Are the data for the study available online?”
55. “Were results generalised to the target population?”

7.3 Checklist items not included in the calculation of score

1. “Time started”
2. “Study authors by surname”
3. “Publication year”
4. “Journal”
5. “Study number”
6. “Number of study hypotheses as specified by authors”
7. “Description of study hypotheses”
8. “Was study hypothesis generating or hypothesis testing?”
9. “Study design as specified by authors”
10. “Target population as specified by authors”
11. “Inclusion / exclusion criteria as specified by authors”
12. “Total number of participants invited to take part in study”
13. “Total number of participants who agreed to take part in study”
14. “Total number of participants who took part in study”

15. “Description of participants included in study”
16. “Were participants recruited solely for the purposes of this study?”
17. “Were participants recruited from a subject pool (potential to be recruited for multiple studies)?”
18. “Estimated number of rounds and questions to be presented (mentioned before results section of paper)?”
19. “Sampling procedure as described by authors”
20. “Use your judgment: was it random or convenience sample?”
21. “Country where study was conducted”
22. “Study start date (/time)”
23. “Study end date (/time)”
24. “Study setting as specified by authors (e.g. lab, online, etc.)”
25. “State what was explicitly randomised (e.g. treatment, question, order, stimuli, etc.)”
26. “Describe the procedure used to generate random assignment sequence”
27. “Describe the method used to conceal the allocation sequence”
28. “If deception was used, how was it described by authors?”
29. “Estimated total number of questions and rounds presented from results section.”
30. “Total number of significance tests reported”
31. “Total number of participants included in analysis”
32. “Total number of participants excluded/dropped/lost to follow up”

33. “Reasons for attrition given by authors”
34. “Were there any missing data?”
35. “If there were missing data, how were they handled?”
36. “Please state any other potential sources of risk of bias”
37. “How were comparison groups they defined? E.g. control, comparison, contrast group.”
38. “If there were baseline differences, please specify what the differences were”
39. “Total number of groups”
40. “What was the unit used in the statistical analyses?”
41. “Were there repeated measures?”
42. “Was a repeated measures analysis used?”
43. “If results were generalised to outsde the target population, how were they jutified by the authors?”
44. “Was a generic generalisation statement made?”
45. “Time ended”
46. “Comments, problems, etc.”

7.4 Checklist results by replication success categories

Table 7.1: Checklist items with 'Yes' or 'No' responses by the replication success categories. Items appear in the. same order as in the original checklist. A dash (-) indicates that the categorical response was not applicable to that item. Items with an asterisk (*) were not included in the calculation of the score. Responses are shown in percentages with rounding error of $\pm 1\%$

Checklist question	Yes	Yes for some	No	Unclear	NA
'Design' section of checklist					
1. Were hypotheses aims or goals clearly stated?					
"Very successful" replicated studies (VS) [$n = 13$]	92	-	8	-	-
"Successful" replicated studies (S) [$n = 13$]	92	-	8	-	-
"Unsuccessful" replicated studies (U) [$n = 13$]	100	-	0	-	-
If study had multiple hypotheses, aims, or goals, were they prioritised?					
VS	15	-	23	23	38
S	0	-	8	38	54
U	31	-	8	15	46
'Participants' section of checklist					
Use your judgment: was sample representative of target population?					
VS	0	-	8	0	92
S	0	-	0	0	100
U	0	-	8	0	92
*Were participants recruited solely for the purposes of this study?					
VS	8	-	8	85	-
S	23	-	0	77	-

U	8	-	23	69	-
5. *Were participants recruited from a subject pool (potential to be recruited for multiple studies)?					
VS	54	-	15	31	-
S	23	-	15	62	-
U	54	-	8	38	-
'Measurement' section of checklist					
Did author(s) explicitly define each outcome variable?					
VS	77	23	0	0	-
S	92	8	0	0	-
U	85	8	0	8	-
Did author(s) explicitly describe how each outcome variable relate to the goals of the study?					
VS	69	31	0	0	-
S	92	8	0	0	-
U	92	8	0	0	-
Did author(s) explicitly explain how each outcome variable was measured?					
VS	77	15	8	0	-
S	92	8	0	0	-
U	92	8	0	0	-

If instruments were used to collect data, did author(s) describe the reliability with regard to the way the instruments are used in a population?					
VS	0	0	100	-	0
S	0	0	100	-	0
U	0	8	92	-	0
10. If an instrument was used to collect data, did author(s) describe the validity with regard to the way the instrument is used in a population?					
VS	8	0	92	-	0
S	0	0	100	-	0
U	0	0	100	-	0
If a physical apparatus was used, did author(s) describe the brand?					
VS	31	-	8	-	62
S	8	-	23	-	69
U	8	-	0	-	92
If a physical apparatus was used, did author(s) describe the model?					
VS	31	-	8	-	62
S	0	-	31	-	69
U	8	-	0	-	92
If a physical apparatus was used, did author(s) describe the design specifications?					
VS	15	-	15	-	69
S	8	-	23	-	69

U	0	-	8	-	92
'Procedure' section of checklist					
Did author(s) report planning sample size in advance?					
VS	0	0	100	0	-
S	0	0	100	0	-
U	0	0	100	0	-
15. Was sample size reported?					
VS	100	-	0	0	-
S	92	-	8	0	-
U	100	-	0	0	-
Did author(s) describe any anticipated sources of attrition due to non-compliance, dropout, death, or other factors?					
VS	8	-	92	0	0
S	0	-	92	8	0
U	8	-	77	15	0
Did author(s) describe how such attrition may affect the generalisability of results?					
VS	0	-	100	0	-
S	0	-	100	0	-
U	0	-	100	0	-
Were pilot study sessions conducted?					
VS	8	-	0	92	-

S	0	-	0	100	-
U	0	-	8	92	-
If pilot sessions were conducted, was the purpose stated?					
VS	0	-	8	0	92
S	0	-	0	0	100
U	0	-	0	0	100
20. Did author(s) describe personnel who collected the data?					
VS	8	-	38	0	54
S	0	-	31	0	69
U	0	-	46	0	54
Did author(s) describe personnel who administered the study?					
VS	0	-	92	0	8
S	0	-	92	0	8
U	8	-	85	0	8
'Allocation and concealment' section of checklist					
Did author(s) explicitly specify randomising treatment to participants?					
VS	31	-	38	8	23
S	46	-	31	0	23
U	46	-	38	0	15
Use your judgment: Was allocation adequately concealed?					
VS	0	-	0	62	38
S	0	-	0	85	15

U	0	-	0	69	31
'Deception' section of checklist					
Were participants informed of the true purpose of the study?					
VS	38	-	23	38	-
S	15	-	15	69	-
U	8	-	54	38	-
25. If deception was used, was it justified by authors?					
VS	0	-	23	-	77
S	8	-	8	-	85
U	0	-	54	-	46
'Boredom' section of checklist					
Was there monetary incentive?					
VS	46	23	23	8	-
S	46	15	23	15	-
U	38	8	38	15	-
'Blinding' section of checklist					
Were responses automatically captured, e.g. computer software programme?					
VS	69	8	15	8	-
S	69	0	8	23	-
U	54	0	15	31	-

If responses NOT automatically captured, was an attempt made to blind personnel collecting data?					
VS	0	-	0	31	69
S	0	-	0	31	69
U	0	-	0	46	54
Was an attempt made to blind participants to treatment group assigned?					
VS	0	-	0	54	46
S	0	-	0	69	31
U	23	-	8	38	31
30. Was an attempt made to blind personnel administering study?					
VS	0	-	0	77	23
S	0	-	0	85	15
U	0	-	0	85	15
Was an attempt made to blind those analysing the data?					
VS	0	-	8	92	-
S	0	-	0	100	-
U	0	-	0	100	-
'Multiple testing' section of checklist					
Were methods used to handle multiple testing?					
VS	0	-	85	15	0
S	8	-	85	8	0
U	8	-	62	23	8

'Attrition' section of checklist					
Did author(s) describe the participants excluded/dropped/lost to follow up					
VS	0	-	15	0	85
S	0	-	8	15	77
U	8	-	38	0	54
Were analyses compared with and without the participants excluded/dropped/lost to follow up?					
VS	0	-	15	0	85
S	8	-	15	0	77
U	0	-	31	15	54
35. Did author(s) describe differences with and without the participants excluded/dropped/lost to follow up?					
VS	0	-	15	0	85
S	0	-	15	8	77
U	0	-	31	8	62
*Were there any missing data?					
VS	15	-	23	62	-
S	15	-	31	54	-
U	15	-	38	46	-
'Comparison group' section of checklist					
Did author(s) report the demography / characteristics of each group?					

VS	23	-	77	-	0
S	23	-	77	-	0
U	8	-	92	-	0
Were groups tested for baseline differences?					
VS	0	-	31	38	31
S	8	-	15	46	31
U	0	-	31	62	8
Was the number of of participants in each group stated?					
VS	92	-	8	0	0
S	77	-	8	15	0
U	54	-	46	0	0
40. Was sample size equally balanced across groups (+/− 10% difference)?					
VS	23	-	23	8	46
S	23	-	23	23	31
U	31	-	15	46	8
'Analysis' section of checklist					
Use your judgment: Was there evidence that data quality was checked (e.g. outliers, illegal values, 'anomalies in the data')?					
VS	23	-	77	-	-
S	15	-	85	-	-
U	15	-	85	0	-

Use your judgment: Was basic data adequately described? (e.g. What was the distribution of the data? Was M +/− SD provided for normally distributed data? Was interquartile range or graphics provided for skewed data?)					
VS	8	-	85	8	-
S	0	-	85	15	-
U	0	-	77	23	-
Was each statistical method described sufficiently to understand what was done? E.g. Spearman's rank correlation and not simply 'correlation'.					
VS	31	23	46	-	-
S	31	23	46	-	-
U	23	62	15	-	-
Did author(s) justify the use of each statistical method?					
VS	0	23	77	0	-
S	0	31	69	0	-
U	0	23	77	0	-
45. Did author(s) describe the underlying assumptions of each analysis?					
VS	0	15	85	-	-
S	0	8	92	-	-
U	0	0	100	-	-
Use your judgment: do underlying assumptions seem reasonable given the data?					

VS	0	-	0	100	-
S	0	-	23	77	-
U	0	-	8	92	-
Did author(s) report examining residuals?					
VS	0	-	92	-	8
S	0	-	62	-	38
U	0	-	85	-	15
Were residuals presented graphically?					
VS	0	-	92	-	8
S	0	-	62	-	38
U	0	-	85	-	15
'Results reporting' section of checklist					
Was a protocol or pre-registered study mentioned?					
VS	0	-	100	-	0
S	0	-	100	-	0
U	0	-	100	-	0
50. Was there any deviation from the protocol/pre-registered study?					
VS	0	-	0	-	100
S	0	-	0	-	100
U	0	-	0	-	100
Use your judgment: were all outcomes tested, reported?					
VS	15	-	31	54	-

S	38	-	15	46	-
U	15	-	23	62	-
Were actual p-values reported (e.g. 0.035 instead of p< 0.05) for each finding except where the p-value is less than 0.001?					
VS	8	46	46	-	0
S	23	46	31	-	0
U	23	38	38	-	0
Were effect sizes presented for each finding?					
VS	8	77	15	-	0
S	31	38	31	-	0
U	38	54	8	-	0
Were interval estimates presented for each finding?					
VS	0	0	100	-	0
S	0	23	77	-	0
U	0	0	100	-	0
55. Were interval estimates presented in each figure, where appropriate?					
VS	54	0	38	-	8
S	31	0	31	-	38
U	15	0	31	-	54
Did the author(s) report the statistical software used?					
VS	15	-	85	-	0
S	8	-	92	-	0

U	8	-	92	-	0
Is the data for the study available online?					
VS	23	-	38	38	-
S	23	-	31	46	-
U	0	-	62	38	-
'Discussion' section of checklist					
58. Were results generalised to the target population?					
VS	8	-	0	-	92
S	0	-	0	-	100
U	0	-	8	-	92

Bibliography

Altman, Douglas G. 1991. *Practical Statistics for Medical Research*. Chapman; Hall.

Altman, Douglas G, and J Martin Bland. 1994a. "Diagnostic Tests 1: Sensitivity and Specificity." *British Medical Journal* 308 (6943): 1552.

———. 1994b. "Diagnostic Tests 2: Predictive Values." *British Medical Journal* 309 (6947): 102.

Altman, Douglas G, Sheila M Gore, Martin J Gardner, and Stuart J Pocock. 1983. "Statistical Guidelines for Contributors to Medical Journals." *British Medical Journal* 286 (6376): 1489–93.

Altman, Douglas G, and David Moher. 2018. "Reply to Letter to the Editor by C. Faggion: Reproducibility and Reporting Guidelines." *Journal of Clincal Epidemiology* 100: 131–32.

Altman, Douglas G, and Iveta Simera. 2016. "A History of the Evolution of Guidelines for Reporting Medical Research: The Long Road to the EQUATOR Network." *Journal of the Royal Society of Medicine* 109 (2): 67–77.

American Psychological Association. 2013. *Publication Manual of the American Psychological Association*. 6th ed.

Anderson, Samantha F, and Scott E Maxwell. 2017. "Addressing the 'Replication Crisis': Using Original Studies to Design Replication Studies with Appropriate Statistical Power." *Multivariate Behavioral Research* 52 (3): 305–24.

APA Publications and Communications Board Working Group on Journal Article Reporting Standards. 2008. "Reporting Standards for Research in Psychology: Why

Do We Need Them? What Might They Be?" *American Psychologist* 63 (9): 839–51.

Austin, Peter C, and Juan Merlo. 2017. "Intermediate and Advanced Topics in Multilevel Logistic Regression Analysis." *Statistics in Medicine* 36 (20): 3257–77.

Begg, Colin, Mildred Cho, Susan Eastwood, Richard Horton, David Moher, Ingram Olkin, Roy Pitkin, et al. 1996. "Improving the Quality of Reporting of Randomised Controlled Trials: The CONSORT Statement." *JAMA* 276 (8): 637–39.

Benjamin, Daniel J, James O Berger, Magnus Johannesson, Brian A Nosek, E-J Wagenmakers, Richard Berk, Kenneth A Bollen, et al. 2018. "Redefine Statistical Significance." *Nature Human Behaviour* 2 (1): 6.

Birnbaum, Michael H, and Jeffrey P Bahra. 2012. "Separating Response Variability from Structural Inconsistency to Test Models of Risky Decision Making." *Judgment and Decision Making* 7 (4): 402–26.

Bland, J Martin, and DG Altman. 1994. "Statistics Notes: One and Two Sided Tests of Significance." *British Medical Journal* 309 (6949): 248.

British Medical Journal. 1996. "The Nuremberg Code (1947)." *British Medical Journal* 313 (7070): 1448. https://doi.org/10.1136/bmj.313.7070.1448.

British Psychological Society. 2010. *Code of Human Research Ethics*. Leicester, UK: British Psychological Society.

Brown, Sacha D, David Furrow, Daniel F Hill, Jonathon C Gable, Liam P Porter, and W Jake Jacobs. 2014. "A Duty to Describe: Better the Devil You Know Than the Devil You Don't." *Perspectives on Psychological Science* 9 (6): 626–40.

Camerer, Colin F, Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, et al. 2016. "Evaluating Replicability of Laboratory Experiments in Economics." *Science* 351 (6280): 1433–6.

Camerer, Colin F, Anna Dreber, Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, et al. 2018. "Evaluating the Replicability of Social Science Experiments in Nature and Science Between 2010 and 2015." *Nature Human Behaviour* 2: 637–44.

Camerer, Colin, Samuel Issacharoff, George Loewenstein, Ted O'Donoghue, and Matthew Rabin. 2003. "Regulation for Conservatives: Behavioural Economics and the Case for 'Asymmetric Paternalism'." *University of Pennsylvania Law Review* 151 (3): 1211–54.

Christensen, Garret, and Edward Miguel. n.d. "Transparency, Reproducibility, and the Credibility of Economics Research." *Journal of Economic Literature*.

Clarke, Geoffrey Mallin, and Robert E Kempson. 1996. *Introduction to the Design and Analysis of Experiments*. Wiley.

Committee on Professional Ethics of the American Statistical Association. 2016. "Ethical Guidelines for Statistical Practice." American Statistical Association.

Cook, Thomas D, Donald Thomas Campbell, and William Shadish. 2002. *Experimental and Quasi-Experimental Designs for Generalised Causal Inference*. Houghton Mifflin Boston.

Cowling, Benjamin J, J Ewart H Shaw, Jane L Hutton, and Anthony G Marson. 2007. "New Statistical Method for Analyzing Time to First Seizure: Example Using Data Comparing Carbamazepine and Valproate Monotherapy." *Epilepsia* 48 (6): 1173–8.

Cox, David Roxbee, and Christl A Donnelly. 2011. *Principles of Applied Statistics*. Cambridge University Press.

Davis, Douglas D, and Charles A Holt. 1993. *Experimental Economics*. Princeton University Press.

Downs, Sara H, and Nick Black. 1998. "The Feasibility of Creating a Checklist for the Assessment of the Methodological Quality Both of Randomised and Non-Randomised Studies of Health Care Interventions." *Journal of Epidemiology & Community Health* 52 (6): 377–84.

Dunbar, George. 2008. *Evaluating Research Methods in Psychology: A Case Study Approach*. John Wiley & Sons.

Ellison, Stephen LR, Vicki J Barwick, and Trevor J Duguid Farrant. 2009. *Practical*

Statistics for the Analytical Scientist: A Bench Guide. Royal Society of Chemistry.

Faralla, Valeria, Francesca Benuzzi, Paolo Frigio Nichelli, and Nicola Dimitri. 2012. "Neuroscience and the Economics of Decision Making." In, edited by Alessandro Innocenti and Angela Sirigu, 146–63. Routledge.

Fletcher, Astrid, Sheila Gore, David Jones, Ray Fitzpatrick, David Spiegelhalter, and David Cox. 1992. "Quality of Life Measures in Health Care. II: Design, Analysis, and Interpretation." *British Medical Journal* 305 (6862): 1145–8.

Flouri, Marilena, Shuyan Zhai, Thomas Mathew, and Ionut Bebu. 2017. "Tolerance Limits and Tolerance Intervals for Ratios of Normal Random Variables Using a Bootstrap Calibration." *Biometrical Journal* 59 (3): 550–66.

Frederick, Shane, George Loewenstein, and Ted O'Donoghue. 2002. "Time Discounting and Time Preference: A Critical Review." *Journal of Economic Literature* 40 (2): 351–401.

Fréchette, Guillaume R, and Andrew Schotter. 2015. *Handbook of Experimental Economic Methodology*. Oxford University Press.

Friedman, Daniel, and Shyam Sunder. 1994. *Experimental Methods: A Primer for Economists*. Cambridge University Press.

Goldstein, Harvey. 1984. "Present Position and Potential Developments: Some Personal Views. Statistics in the Social Sciences." *Journal of the Royal Statistical Society. Series A* 147 (2): 260–67.

Goodman, Steven N, Daniele Fanelli, and John PA Ioannidis. 2016. "What Does Research Reproducibility Mean?" *Science Translational Medicine* 8 (341): 341ps12.

Gore, Sheila M, Ian G Jones, and Eilif C Rytter. 1977. "Misuse of Statistical Methods: Critical Assessment of Articles in BMJ from January to March 1976." *British Medical Journal* 1 (6053): 85–87.

Halpern, David. 2015. "The Rise of Psychology in Policy: The UK's de Facto Council of Psychological Science Advisers." *Perspectives on Psychological Science* 10 (6): 768–71.

Han, Ruokang, and Taiki Takahashi. 2012. "Psychophysics of Time Perception and Valuation in Temporal Discounting of Gain and Loss." *Physica A: Statistical Mechanics and Its Applications* 391 (24): 6568–76.

Hand, David J. 2004. *Measurement Theory and Practice*. Hodder Arnold.

Hardisty, David J, and Elke U Weber. 2009. "Discounting Future Green: Money Versus the Environment." *Journal of Experimental Psychology: General* 138 (3): 329–40.

Hardisty, David, Katherine F. Thompson, David Krantz, and Elke U. Weber. 2013. "How to Measure Time Preferences: An Experimental Comparison of Three Methods." *Judgment and Decision Making* 8 (3): 236–49.

Henrich, Joseph, Steven J Heine, and Ara Norenzayan. 2010a. "Most People Are Not WEIRD." *Nature* 466 (7302): 29.

———. 2010b. "The Weirdest People in the World?" *Behavioural and Brain Sciences* 33 (2-3): 61–83.

Hertwig, Ralph, and Andreas Ortmann. 2001. "Experimental Practices in Economics: A Methodological Challenge for Psychologists?" *Behavioural and Brain Sciences* 24 (3): 383–403.

Higgins, Julian, and Douglas G Altman. 2008. "Assessing Risk of Bias in Included Studies." In *Cochrane Handbook for Systematic Reviews of Interventions*, edited by Julian Higgins and Sally Green, 187–241. Cochrane handbook for systematic reviews of interventions.

Hutton, Jane L. 1995. "Statistics Is Essential for Professional Ethics." *Journal of Applied Philosophy* 12 (3): 253–61.

Imai, Taisuke, Tom Rutter, and Colin Camerer. 2019. "Meta-Analysis of Present-Bias Estimation Using Convex Time Budgets." *MetaArXiv (Unpublished)*.

Ioannidis, John PA, Tom D Stanley, and Hristos Doucouliagos. 2017. "The Power of Bias in Economics Research." *The Economic Journal* 127 (605): F236–F265.

Jüni, Peter, Douglas G Altman, and Matthias Egger. 2001. "Assessing the Quality

of Controlled Clinical Trials." *British Medical Journal* 323 (7303): 42–46.

Kagel, John H, and Alvin E Roth. 1997. *The Handbook of Experimental Economics*. Princeton University Press.

Kessler, Judd B, and Alvin E Roth. 2012. "Organ Allocation Policy and the Decision to Donate." *American Economic Review* 102 (5): 2018–47.

Kilkenny, Carol, Nick Parsons, Ed Kadyszewski, Michael FW Festing, Innes C Cuthill, Derek Fry, Jane Hutton, and Douglas G Altman. 2009. "Survey of the Quality of Experimental Design, Statistical Analysis and Reporting of Research Using Animals." *PLOS ONE* 4 (11): e7824.

Lempert, Karolina M, Paul W Glimcher, and Elizabeth A Phelps. 2015. "Emotional Arousal and Discount Rate in Intertemporal Choice Are Reference Dependent." *Journal of Experimental Psychology: General* 144 (2): 366–73.

Lim, K. T. K. 2018. "Microthesis: Statistical Practice and Replication Success in Behavioural Science." *London Mathematical Society Newsletter*, no. 477: 31–32.

Loewenstein, George, and Richard H Thaler. 1989. "Anomalies: Intertemporal Choice." *Journal of Economic Perspectives* 3 (4): 181–93.

Loken, Eric, and Andrew Gelman. 2017. "Measurement Error and the Replication Crisis." *Science* 355 (6325): 584–85.

Makel, Matthew C, Jonathan A Plucker, and Boyd Hegarty. 2012. "Replications in Psychology Research: How Often Do They Really Occur?" *Perspectives on Psychological Science* 7 (6): 537–42.

Mazur, James E. 1987. "Quantitative Analyses of Behaviour." In, edited by Mazure Commons M. L., 5:55–73. Hillsdale.

McKerchar, Todd L, Stephen Pickford, and Shannon E Robertson. 2013. "Hyperboloid Discounting of Delayed Outcomes: Magnitude Effects and the Gain-Loss Asymmetry." *The Psychological Record* 63 (3): 441–51.

Munafò, Marcus R, Brian A Nosek, Dorothy VM Bishop, Katherine S Button, Christopher D Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wa-

genmakers, Jennifer J Ware, and John PA Ioannidis. 2017. "A Manifesto for Reproducible Science." *Nature Human Behaviour* 1 (0021).

Needleman, Ian G. 2002. "A Guide to Systematic Reviews." *Journal of Clinical Periodontology* 29 (s3): 6–9.

Nichols, Thomas E, Samir Das, Simon B Eickhoff, Alan C Evans, Tristan Glatard, Michael Hanke, Nikolaus Kriegeskorte, et al. 2017. "Best Practices in Data Analysis and Sharing in Neuroimaging Using MRI." *Nature Neuroscience* 20 (3): 299–303.

Open Science Collaboration. 2015. "Estimating the Reproducibility of Psychological Science." *Science* 349 (6251): aac4716.

Peng, Roger. 2015. "The Reproducibility Crisis in Science: A Statistical Counterattack." *Significance* 12 (3): 30–32.

Read, Daniel. 2004. "Intertemporal Choice." In *Blackwell Handbook of Judgment and Decision Making*, edited by Derek J Koehler and Nigel Harvey, 424–43. Wiley Online Library.

Siddaway, Andy P, Alex M Wood, and Larry V Hedges. 2018. "How to Do a Systematic Review: A Best Practice Guide for Conducting and Reporting Narrative Reviews, Meta-Analyses, and Meta-Syntheses." *Annual Review of Psychology*, no. 0.

Simmons, Joseph P, Leif D Nelson, and Uri Simonsohn. 2011. "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22 (11): 1359–66.

Smith, Vernon L. 1994. "Economics in the Laboratory." *Journal of Economic Perspectives* 8 (1): 113–31.

Spellman, Barbara A. 2015. "A Short (Personal) Future History of Revolution 2.0." *Perspectives on Psychological Science* 10 (6): 886–99.

Thaler, Richard. 1981. "Some Empirical Evidence on Dynamic Inconsistency." *Economics Letters* 8 (3): 201–7.

Thompson, Simon G, and Julian Higgins. 2002. "How Should Meta-Regression Analyses Be Undertaken and Interpreted?" *Statistics in Medicine* 21 (11): 1559–73.

Turner, Lucy, Larissa Shamseer, Douglas G Altman, Kenneth F Schulz, and David Moher. 2012. "Does Use of the CONSORT Statement Impact the Completeness of Reporting of Randomised Controlled Trials Published in Medical Journals? A Cochrane Review." *Systematic Reviews* 1 (60): 1–7.

Wasserstein, Ronald L, Nicole A Lazar, and others. 2016. "The ASA's Statement on *p*-Values: Context, Process, and Purpose." *The American Statistician* 70 (2): 129–33.

Wicherts, Jelte M, Coosje LS Veldkamp, Hilde EM Augusteijn, Marjan Bakker, Robbie Van Aert, and Marcel ALM Van Assen. 2016. "Degrees of Freedom in Planning, Running, Analysing, and Reporting Psychological Studies: A Checklist to Avoid *p*-Hacking." *Frontiers in Psychology* 7: 1832.

Wilkinson, Leland, and the Task Force on Statistical Inference. 1999. "Statistical Methods in Psychology Journals: Guidelines and Explanations." *American Psychologist* 54 (8): 594–604.

Xu, Lijuan, Zhu-Yuan Liang, Kun Wang, Shu Li, and Tianzi Jiang. 2009. "Neural Mechanism of Intertemporal Choice: From Discounting Future Gains to Future Losses." *Brain Research* 1261: 65–74.